

# **PROJECT: Analysis of United States National Occupational Employment and Wage data**

**And exploring the correlation between crime rate in US states and the income level in those states**

JYOTI CHAUDHARY

DATA WRANGLING AND HUSBANDRY, RUTGERS UNIVERSITY

Email: [jc2180@rutgers.edu](mailto:jc2180@rutgers.edu)

My project focuses on analyzing and making inferences on “**United States National Occupational Employment and Wage Estimates**”.

- I have identified the top paying and lowest paying occupations within US. Plotted them via boxplot.
- Implemented R code to plot the density graph that shows, via a plot for each major occupation type, the count of employees that fall in various salary range.
- Extracted 2015 US Crime data from FBI website. This data lists the crime statistics by state.
- Coded a Shiny app that allows the selections of ‘occupation type’ and the ‘Map type’ (Annual Wage or Location Quotient). If annual Wage is selected as map type, the app plots a histogram that shows the state wise employment count for the selected occupation. Another choropleth is plotted to show the Annual Mean wage by state for the selected occupation. If ‘Location

Quotient', the histogram will still plot the same data , however, the choropleth map will show the location quotient by state for the selected occupation. The higher location quotient in a state indicates that the state has higher share of occupation's employment than the US as a whole.

- Another dropdown on shiny app to select the type of Crime (robbery, assault... ). Another choropleth is plotted to show the crime rate (per 100,000) by state for the selected crime type.
- Correlation identified between the income level and the crime rate in various states as detailed later in this report.

## Data

The data for this project has been downloaded from “Bureau Of Labor Statistics” website - <https://www.bls.gov/bls/blswage.htm>

And the Crime rate data from FBI website - (<https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015>)

National Employment and Wage Estimates – “National\_salary.xls”

State wise Employment and Wage Estimates – “state\_May2015\_d1”

2015 US Crime Rate – “US\_Crime\_2015”

**The Employment and Wage estimates data is for the following major occupational groups:**

- Management Occupations
- Business and Financial Operations Occupations
- Computer and Mathematical Occupations

- Architecture and Engineering Occupations
- Life, Physical, and Social Science Occupations
- Community and Social Service Occupations
- Legal Occupations
- Education, Training, and Library Occupations
- Arts, Design, Entertainment, Sports, and Media Occupations
- Healthcare Practitioners and Technical Occupations
- Healthcare Support Occupations
- Protective Service Occupations
- Food Preparation and Serving Related Occupations
- Building and Grounds Cleaning and Maintenance Occupations
- Personal Care and Service Occupations
- Sales and Related Occupations
- Office and Administrative Support Occupations
- Farming, Fishing, and Forestry Occupations
- Construction and Extraction Occupations
- Installation, Maintenance, and Repair Occupations
- Production Occupations
- Transportation and Material Moving Occupations

Columns read from the National and State data files to draw inferences:

- Occupation Title

- Employment Count
- Annual Mean Wage
- Annual 10th percentile wage, 25th percentile, median (50<sup>th</sup> percentile) wage, 75th percentile wage, 90<sup>th</sup> percentile wage
- Level – Major , Detail
  - Major – Computer & Mathematical Occupations
  - Detail –
    - Computer and Information Research Scientists
    - Information Security Analyst
    - Web Developer
    - Database Administrator

Additional Columns read from State wise data

- Location Quotient
  - (The location quotient represents the ratio of an occupation's share of employment in a given area to that occupation's share of employment in the U.S. as a whole. )

**The Crime rate data is for year 2015 and is for the following crime categories. All analysis is done on the “rate per 100,000” data for each crime category.**

Violent crime  
Murder and nonnegligent manslaughter  
Rape  
Robbery  
Aggravated assault  
Property crime  
Burglary  
Larceny-theft  
Motor vehicle theft

### **Steps involved in data cleaning**

Crime data – “US\_Crime\_2015”

- Extracted the data file. For each state, the file had 3 rows for – “2014, 2015 and ‘Percent Change’”. I decided to extract only the rows for 2015. However, the ‘Area’ column had state name only for the first row which was for ‘2014’. Coded the ‘For loop’ to copy the state name into ‘2015’ row from previous row (which was for ‘2014’). Code snippet below (in red):

```
for(i in 2:nrow(crime_file)) # FOR loop to fill Area column value correctly wherever its blank or
NA in the dataset
{
  if (is.na(prevarea[i])) {
    prevarea[i] <- prevarea[i-1] # when area value blank, fill it from previous row area value
  }
}
```

- Extracted the rows for year '2015' and edited column names on the dataframe.
- The 'Area' column (renamed to 'State' in R dataframe) had data such as below with 'commas' and 'digits' appended at the end and also prefixed with spaces.

United States Total6, 7, 8, 9

Northeast6

Maine6

New York6

Used below R code to clean the column data (highlighted in yellow)

```
Crime_file1$State <- str_trim(gsub(",", "", str_trim(str_replace_all(Crime_file1$State, "[0-9]+", ""))))
```

- As you can see from the above list of 'Area' values, some of the rows are for regions (not the US State) - United States Total, Northeast, Midwest. Used Left join to fix this. Created a dataframe with US state codes and state names. Joined this dataframe with the Crime\_file dataframe to extract only valid state rows.

```
crime_file3 <- left_join(Crime_file1, state_list, by=c("State" = "state")) # join crime table and state list table on state name
```

```
crime_file4 <- crime_file3 %>% filter(!is.na(state.code)) # select rows that have valid value for state code
crime_file4$State <- str_trim(crime_file4$State)
```

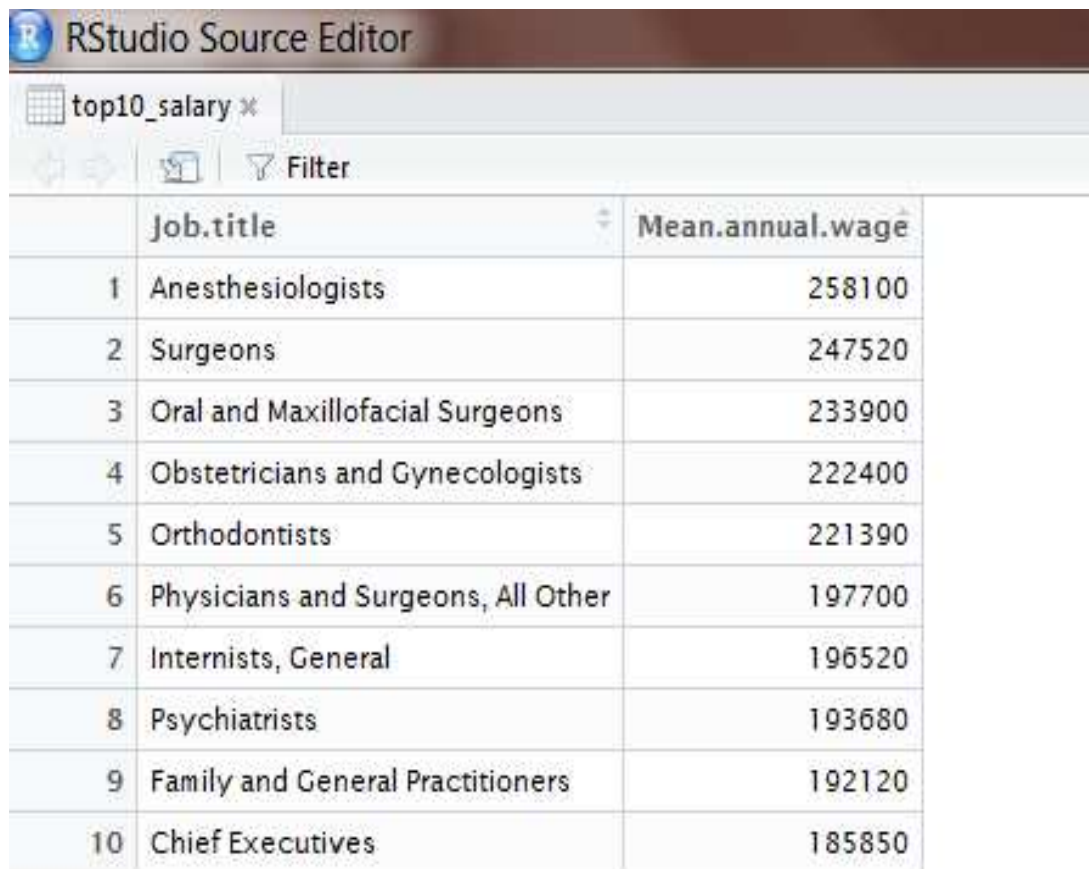
- The datasheet, when loaded in R, all columns were of type 'characters' and converted counts to have decimal and trailing zeros (as "56171281.000000"). Used 'round' and 'as.numeric' to fix this.

- Added a new column to dataframe to hold the total crime count per state for all crime categories.

```
crime_file4$Total.rate <- rowSums(crime_file4[,c(6,8,10,14,16,18,20,22,24)]) # Total the "crime
rate per 100,000" in a new column 'Total.rate'
```

## Graphs / Plots generated from the data

### R code to list the national top paying occupations



The screenshot shows the RStudio Source Editor with a file named 'top10\_salary'. The table displayed has two columns: 'Job.title' and 'Mean.annual.wage'. The data is as follows:

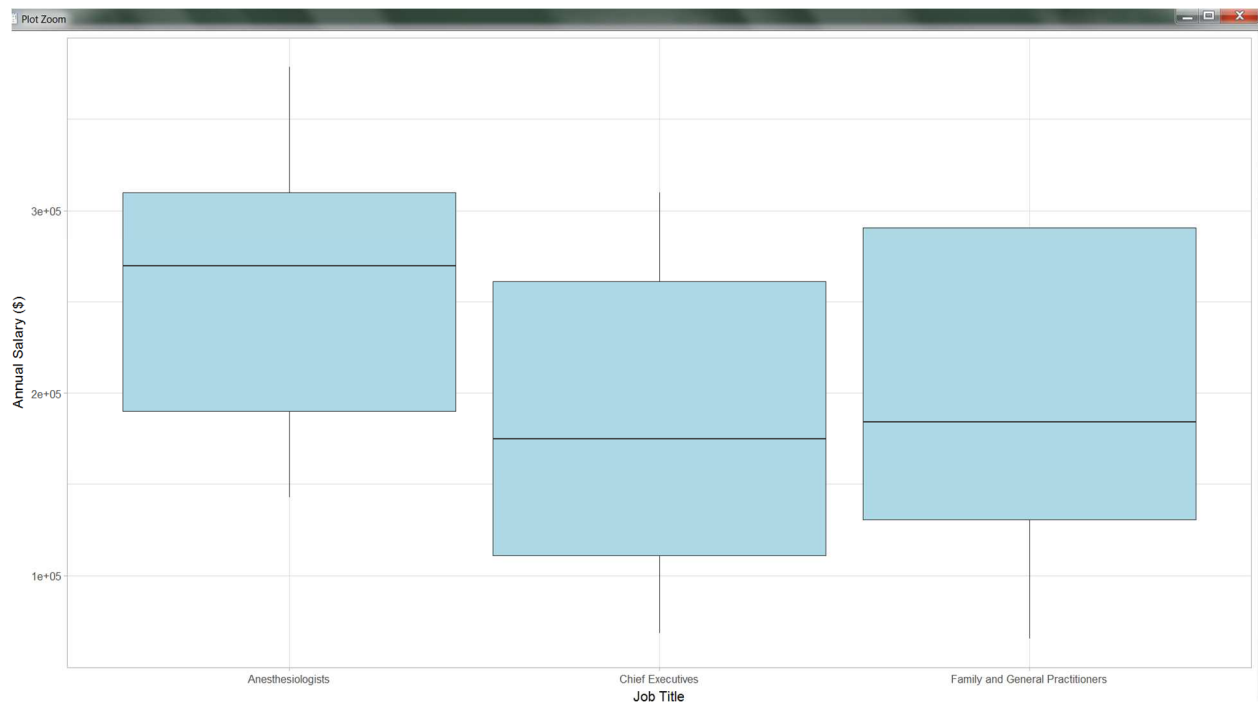
	Job.title	Mean.annual.wage
1	Anesthesiologists	258100
2	Surgeons	247520
3	Oral and Maxillofacial Surgeons	233900
4	Obstetricians and Gynecologists	222400
5	Orthodontists	221390
6	Physicians and Surgeons, All Other	197700
7	Internists, General	196520
8	Psychiatrists	193680
9	Family and General Practitioners	192120
10	Chief Executives	185850

### R code to list the national lowest paying occupations

RStudio Source Editor		
low10_salary x		
Filter		
	Job.title	Mean.annual.wage
1	Dining Room and Cafeteria Attendants and Bartender ...	21400
2	Amusement and Recreation Attendants	21360
3	Ushers, Lobby Attendants, and Ticket Takers	21060
4	Cashiers	20990
5	Counter Attendants, Cafeteria, Food Concession, an...	20590
6	Hosts and Hostesses, Restaurant, Lounge, and Coffe...	20530
7	Dishwashers	20360
8	Shampooers	20350
9	Combined Food Preparation and Serving Workers, Inc...	19710
10	Cooks, Fast Food	19610

Box Plots to show Annual mean wage of 3 top paying occupations (Anesthesiologist, Chief Executives, Family and general practitioners) . The box plots are plotted with considering 10th percentile wage, 25th percentile, median (50<sup>th</sup> percentile) wage, 75th percentile wage and the 90<sup>th</sup> percentile wage.



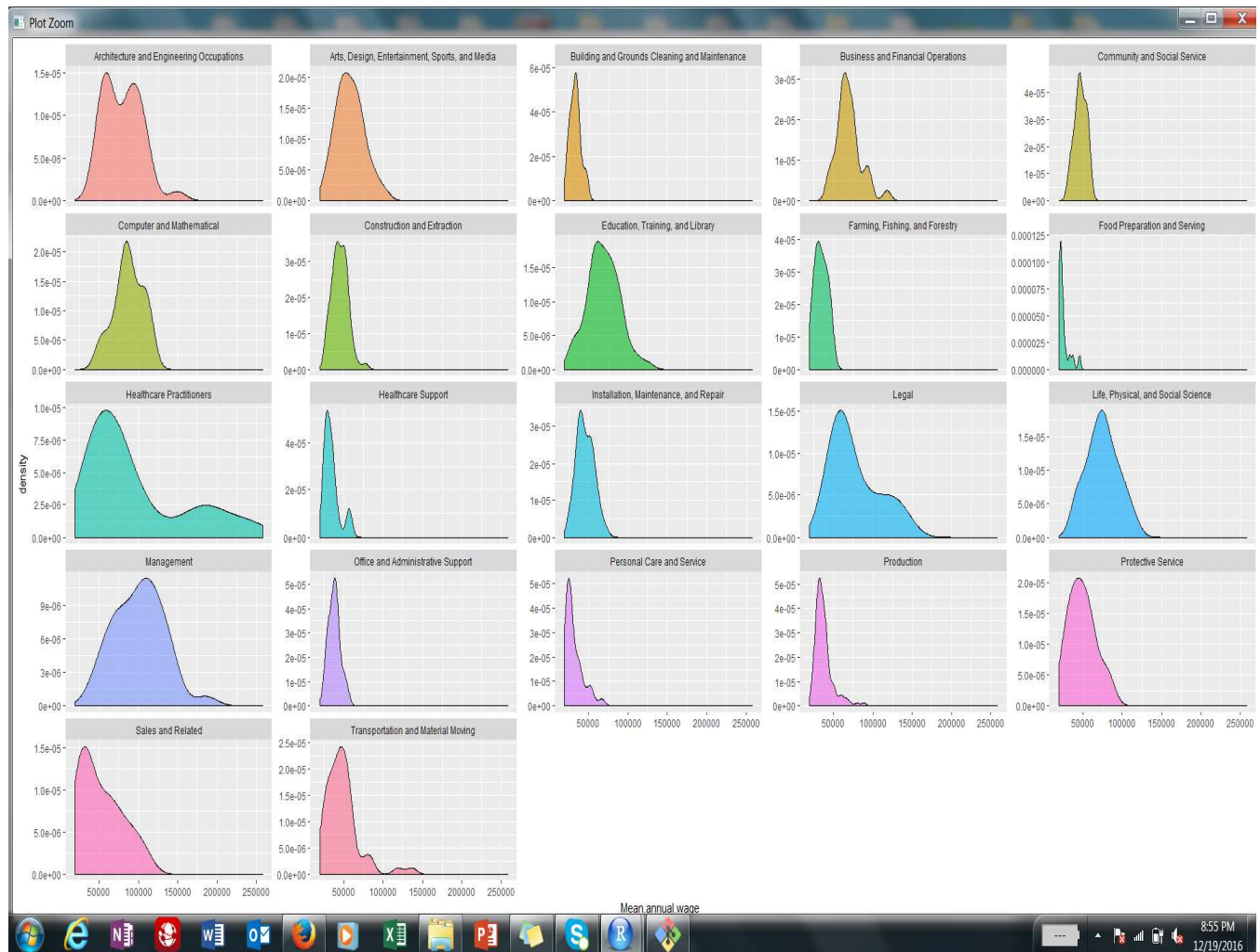


## Box Plot for lowest paying occupations



Density graph for Major Occupations (based on National data).

Salary range plotted on X-axis and Employment count plotted on Y-axis.



**Some observations made from density graph:**

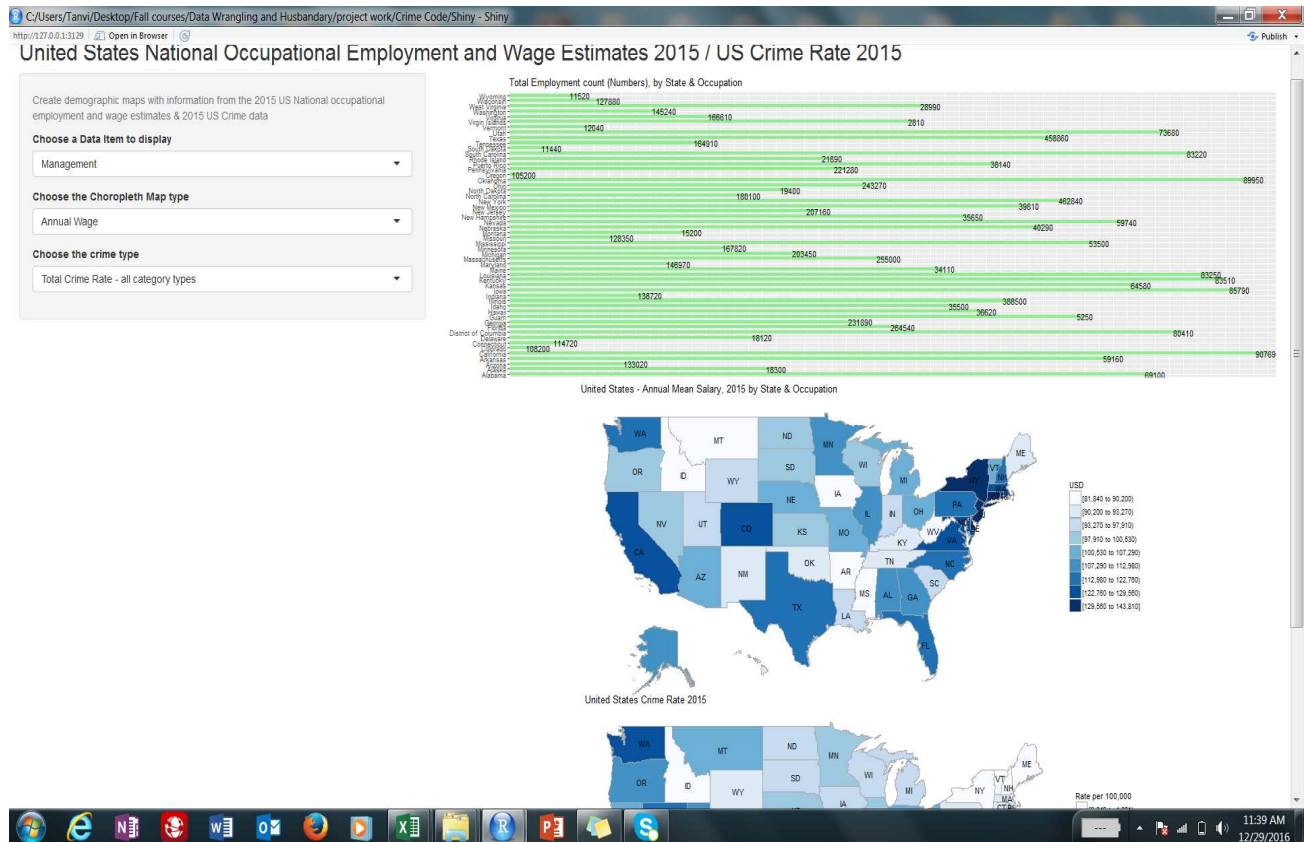
- a) "Healthcare Practitioners" occupation have certain practitioners paid high salaries in the range > 250K-300K.

- b) Most practitioners in "Farming, Fishing and Forestry" are paid less than < 50K
- c) Most practitioners in "Transportation and Material Moving" occupations are paid low salaries < 50K. However, a small percentage of practitioners fall within high salary range >100K (such as Aircraft Pilots and Flight Engineers)

### **SHINY APP**

A "Shiny app" designed to -

- a) Select the Occupation type
- b) Select the Choropleth map type – "Annual Wage or Location Quotient"
- c) Select the Crime Category
- d) Plot the Horizontal bar graph to plot the Employment count by State & Occupation
- e) Plot the State Choropleth map for Annual Mean Salary/ Location Quotient by State and Occupation
- f) Plot the State Choropleth map for crime rate by State and Crime category

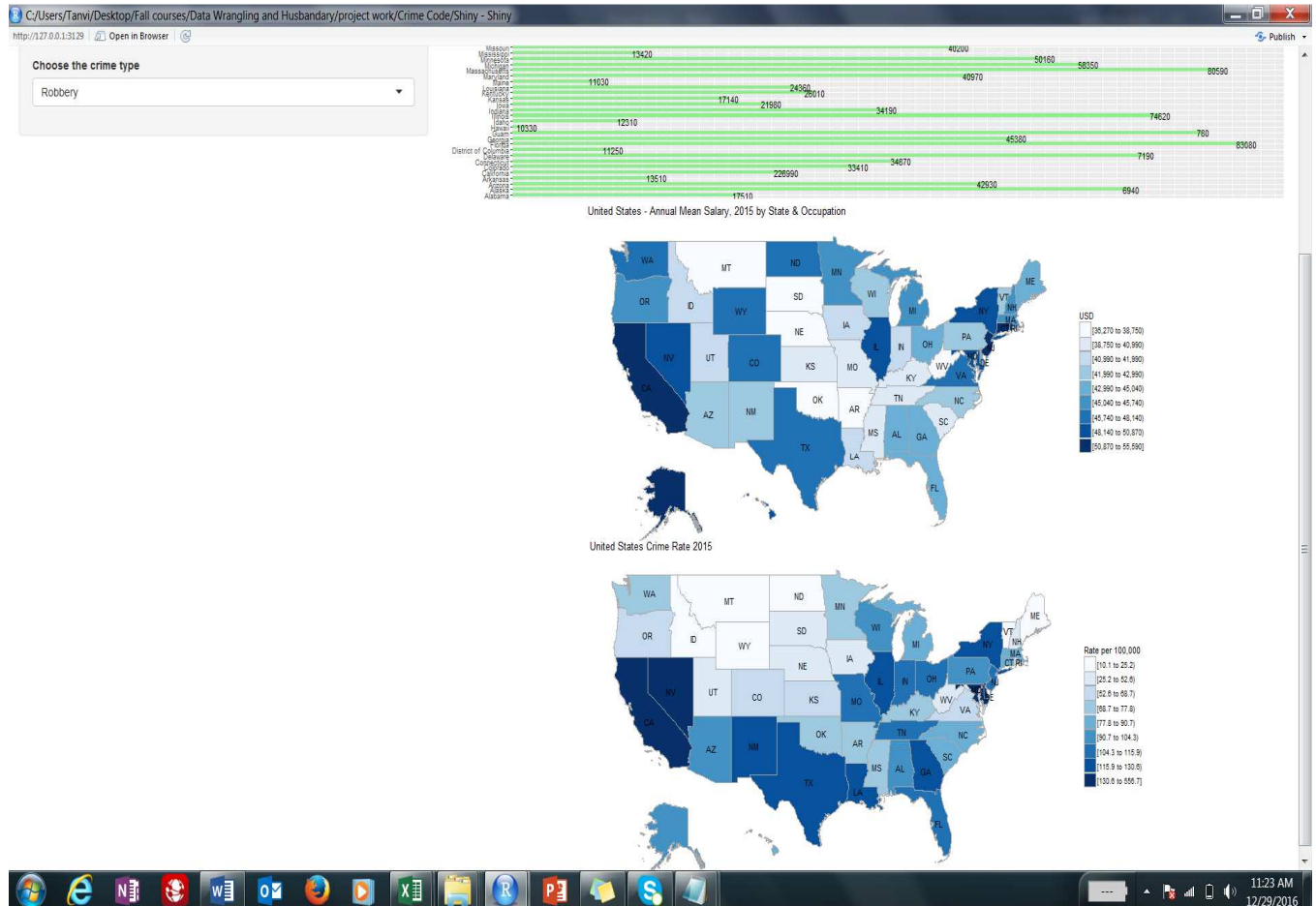


## Inferences from the Plots

California (CA) & Washington (WA) has high concentration of top paying jobs (Computer and Mathematical, Architecture & Engineering, Management)

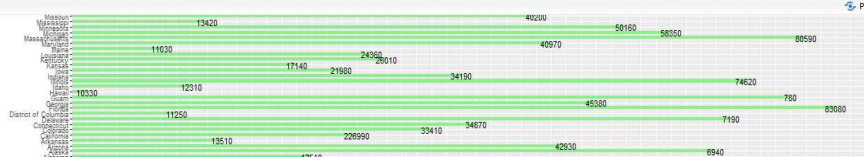
Nevada (NV) and Montana (MT) has high concentration of low paying jobs (Community and Social Service, Food preparation and serving, Building and Grounds cleaning)

CA has high income level and NV has low income level. Both states are adjacent. The Crime rate for 'Robbery' and 'Vehicle Theft' are highest in these two states.

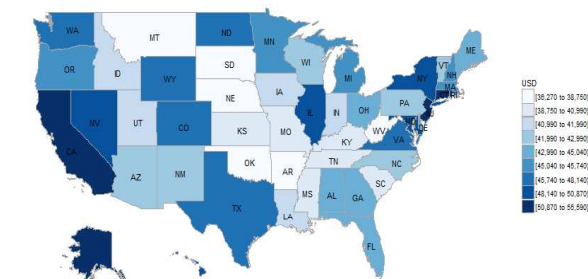


Choose the crime type

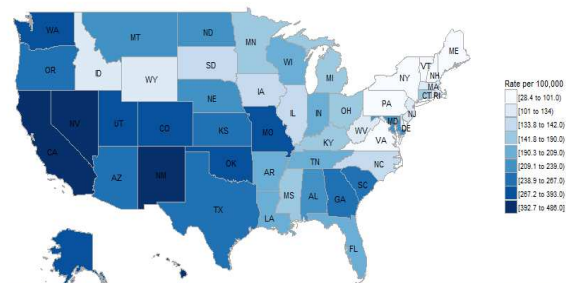
### Vehicle Theft



United States - Annual Mean Salary, 2015 by State &amp; Occupation



United States Crime Rate 2015



Washington with high income jobs and have high rate of 'Property Crime' and 'Burglary'. Nevada with low income has high rate of 'Burglary'.

