2017.9

# CSE7331 DATA MINING

## PROJECT 1

SHIYANG ZHANG
SOUTHERN METHODIST UNIVERSITY
SMU ID: 47319809

## Contens

# 0.Executive Summary:

We are interested in how federal, non-military employment was impacted by the transition from president George Bush to president Barack Obama to learn about the impact of the difference in policy between the two administrations on the federal government. In this project, I use data from 2001 to 2014 to see the change of federal employeement in different years. I mainly choose the first quarter of year 2001, 2008 and 2014. Reasons are listed below:

- The same quarter of a year can decrease error when comparing.
- 2001 is the first president year of Bush, 2008 is the last year of Bush. And 2014 is the most recent data that from Obama. Also, 2008 near the middle of 2001 and 2014.
- For data about 2015 and 2016, the structure of data are quite different from data of 2001 to 2014. So I use data in 2014 instead of data in 2016.

At the beginning, I analyze the business understanding for president Bush and president Obama, one is Republic, one is Demostic. Therefore, many of their policies are different.
Then I describe variables inside of source data files, including their data type and introduce each of them briefly. In addition, I analyze data quality like missing data, duplicated data and ouliers. Then using code to handle those quality problems. Then I do the statistic analyzation for each variable.
At next part, I choose five most important varibles to visualize and analyze. Most cases I compare these variables between diferent years to see the where the change happens.
Next, I explore relationships between variables, like pay and education. Using plot, baxplot and correlation function, I conclude that some variables are positive related to another attribute.
Finally, I consider the method to analyze changes over time. This is a real realistic problem that happens in our daily work and study.

**In conclusion, analyzing nominal variable, counting is a good idea. As for interval/order variable, boxplot is a good method. For ratio variable, histogram is good for use.**
**Agencies pay more as time grows.**
**Pay has positive correlation with LOS.**
**Education and pay have positive correlation with each other.**
**Age between 50-54 earn most salary. Old person become earn more than before.**
**People's education level become higher.**
**In all group of LOS, their 1$^{st}$ and 3$^{rd}$ percentile of education level are similar.**

The report has been complied according to the guidelines of CRISP-DM Framework.

# 1.Bussiness understanding:

The employeement changes are the most important thing I want to research at this project. So main actors are president Bush(2001) and president Obama(2008). President Bush is republic while Obama is demostic. They are some changes in different fields.

| | BUSH | OBAMA | Influences on employment |
|---|---|---|---|
| **Abortion is a woman's unrestricted right** | Disagrees | Strongly Agrees | Hire more female as employee |
| **Legally require hiring women & minorities** | Disagrees | Strongly Agrees | Hire more female and minorities as employee |
| **Comfortable with same-sex marriage** | Strongly Disagrees | Agrees | Homosexual people feel happy than before, work better and harder. Those people easy to find job than before |
| **Make voter registration easier** | Disagrees | Strongly Agrees | Employer hiring people without such worries. People in different job are equal to vote |
| **Stricter punishment reduces crime** | Strongly Agrees | Disagrees | Crime probabiliry may increase. Security apartment need more people |
| **Expand ObamaCare** | Disagrees | Strongly Agrees | Insurance company need more employees |
| **Higher taxes on the wealthy** | Strongly Disagrees | Strongly Agrees | Money hands out to employees become more intensive, decrease the difference between rich and poor |
| **Privatize Social Security** | Strongly Agrees | Strongly Disagrees | Strong information security of every agency. |
| **Expand the military** | Strongly Agrees | Disagrees | Military related agency will curtail employee number |
| **Avoid foreign entanglements** | Disagrees | Agrees | People in USA focus more on domestic affairs |

Table 1. Different attitude to topics

Some important incidents:

- Affordable Care Act, Obama care in another word, was approved bby the Democratic Party in 2008. This thing triggered insurance company need more employees.
- 2001 9.11 incident: security apartment become stressed out than ever before. Those agencies may hire more people to do security job. Government employees handles lots of affairs like hand out pension, counting death people list, etc. They will hire more people doing this. Some government people died in that accident. So government hire new people after that.
- Afganitan Government take over the task to defeat Taliban from United States. US army withdraw from Iraq. So the military related agency will curtail.

# 2. Describe data:

## 2.1 Data from data file:

- Pseudo ID: This is ID number for employees, and it is also a nominal attribute. Everyone's ID should be unique. It consists of numbers.
- Name: This is employee's name and a nominal attribute. I think its value can sometimes be the same.
- Date: This is the date that these data were reported. It is an interval attribute. There is only one date for each file.
- Agency: This is some agencies that employees worked at. It uses some abbreviation to represent each agency. This is also a nominal attribute.
- Station: It indicates locations of employees. Value consists of numbers. Codes are composed of State code (pos 1-2), City code(pos3-6), and County code(pos7-9). So these numbers are also nominal attribute.
- Age: This is age range of each employee. This is order attribute. Most people's age are between 30-55.
- Education: It uses some numbers to represent different education level. Difference between adjacent level is 1. The number is higher, the education level is higher. So I think it is an order attribute. For example, most people are 13--bachelor degree.
- Pay plan: It uses two capital letters to represent different pay plan. It explains for what reason or which category money were given out. This helps companies clear about where did they spend money, so that they can control budget better. Since there is no meaning to compare two different values. It is a nominal attribute.
- Grade: The grade used to determine an employee's basic pay rate. It is an order attribute.
- Los: Length of service represents for how long this person work in this agency. It is a order attribute.
- Occupation: it is a nominal attribute, represent people's job.
- Category: It consists of 6 types—ABCOPT. There are administrative, blue collar, clerical, other white collar, professional and technical. Maybe it divides different occupation into 6 categories. I think this is a nominal attribute.
- Pay: This represents each employee's salary. It is a ratio attribute.
- Supervisory Status: It uses numbers represent different position, like manager, supervisor, team leader, etc. Most people are 'other positions'. It looks like an interval or ratio attribute. But it is meaningless to do mean or median on those numbers. So it is a nominal attribute.
- Appointment: Different career plan represented by numbers. It is a nominal attribute.
- Schedule: Different work schedule represented by letters. Like part time, full time…It is a nominal attribute.
- NSFTP: It is also a career plan for employees. Only two types, 1 and 2.

## 2.2 Data from agency file:

Agency ID: Represented by numbers, it is a nominal attribute.
Agency Name: Represented by strings, it is a nominal attribute.

# 3.Data quality:

## 3.1 Missing Values:

There are different kinds of missing values for variables. For example, for variable 'station', some values are '#########', which are useless and countless. For variable 'age', 'UNSP' is also missing values. Meanwhile, some variables have "" and "*" values. All of those are massing values. I use 'NA' to replace all of them. So when we do summary, we will know 'NA' represent missing values.

## 3.2 Duplicate data:

When we do summary function of those variables, it is very clear to see some duplicate data. For example, employee ID is supposed to be the unique variable. But some ID numbers are shown more than once. However, after comparing other variables, we can understand why some ID are duplicated. Maybe this person worked in one agency at different occupation. Or maybe he changed his job to another agency. So we cannot just delete this row. For people worked in one agency within a quarter, I use the data with high salary and delete information of lower one. Also, if a person has multiple records with different age, I will choose the largest age as his record. For two records are complete the same, I think it is good to delete one of them.

Form the file Status_Non_Dod_2001_03, a person named Cerqueira, Manuel D has three repeated ID records, a very interesting thing is that at the first quarter of 2001, his LOS are '1-2','5-9','5-9'. I do not think that in one quarter person's LOS will increase so much. And correspondingly, his pays are increasing from 1114 to 39131, finally 117600. So I think this is a mistake.

## 3.3 Outliers:

There are some outliers exist in different data file. I do not think delete them is the best idea. Because some of them are useful when doing analyzing. There are some methods to deal wth that. We can separate those data into different groups by numeric value. Delete the largest group and smallest group. This is helpful doing mathmetic analyzation. Also, we can use median instead of mean at some point.

# 4.Statistic analyzation for variables

In order to describe statistic property for each variable, I use the data from "Status_Non_Dod_2001_03.txt" as an example to introduce. As for the mode of each variable, all of them are "numeric".

PseudoID:
It is a nominal variable.The mode of ID is numeric. The contingency of two
identical ID should be 0 if it indicate the same person. The correlation with other attributes should be 0 since the numeric transformation of ID will not influence other attribute's value.

Name:
Name is also a nominal attribute. Its mode is numeric. Also, it has no correlation with other variables.

Date:
There is only one data here: 20010331:1081445.

Agency:
We need to count how many people worked in this agency this time. The agency with most employees are VATA(VETERANS HEALTH ADMINISTRATION) hiring 204634 people at this quarter. Its hiring amount is double to the second agency. This agency must be very important for US government.

Station:
This attribute has lots of NA(invalid value). So many people 's working station are not recorded. The places are relatively scattered. Not concentrate on one place except for top 5 popular places. The reason can be this city is a business metropolis, like NY. Or a political center, like Washington DC. Or maybe this city is so beautiful with a large of population that government needs to expand enrollment, like LA.

Age:
Values for Age are range instead of specific value. So it is inconvenient to do the geometric mean, it's better do the counting. Most people are 45-49.

Education:
The minimum education value is 1. The maximum value is 22. The first quarter is 6. The third quarter is 13. Median is 10. Mean is 10.17. Median and mean are close that no special outliers. The number 6 represent completion of terminal occupational program. 13 represents bachelor's degree. So employees in government are not very high educated. Most people's jobs are not technology related or do not need strong knowledge about some specific area. The enrollment training will teach something that are easy to learn.
Pay plan:
Most people are in pay plan GS, which is general schedule.

Grade:
The number of NA is large in this attribute. After accounting the numeric property, we know that minimum grade is 0. Maximum grade is 97. First quarter is 5. Third quarter is 12. Median is 9. Mean is 8.91. I searched that 97 is not an outlier since many people's grades are larger than 90. But the mean is so small. So most people have grade between 5-12. Most people's grades are 10,11,12.

LOS:
The value is a range, so we count numbers to analyze this attribute. In government agency, most people work 10-14 years. Then are 15-19 years.

Occupation:
Most people are in occupation 0303. For top 10 or more occupation, I think it should be very common for each agency and need large amount people doing this job together. Almost each agency has this occupation like accountant, security, etc.

Category:
Most people are A, it's administrative, then P--professional. Then technical. Minimum people are white collar.

Pay:
The minimum pay is 0. First quarter is 34151. Median is 48827. Third quarter is 69206. Mean is 53796. Maximum is 230628. It shows that most people's salaries are between 34151 to 69206. Mean is bigger than median shows that many people's salary are much higher than commons.

Supervisory Status:
Most people are 8—other positions. The second is 2—supervisor or manager. This attribute is nominal without any order relation. Because that the order of amount of people are not the same with order of value.

Appointment:
Count for each value. Most is value 10: 671518.

Schedule:
Count for different work schedule. Most is F: 956766.

NSFTP:
Count for full time or seasonal. Most people(914835) are Non-Seasonal, Full-Time, Permanent Employees. Other people(166610) are Not Non-Seasonal, Full-Time, Permanent Employees

# 5. Variable visualization

## 5.1 Age:

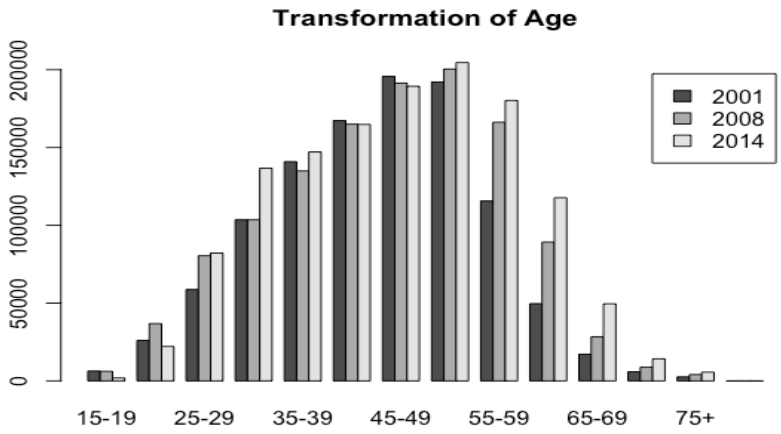


Figure 1. Employee's Age Transformation

In Figure 1, x axis represents ages of employees, y axis represents amount of people of each group. And I also make a comparison between 2001,2008 and 2014. From figure 1, we can see that most people are age 45-55 in all three years. For 2001, most employees are age 45-50, but for 2008 and 2014, most employees are age 50-55. Therefore, the transformation can be regarded that people start working and retire later than before. The reason maybe that people regard diploma as more important than ever. So most people start working from 25. Another reason is that government like experienced people than raw recruit. So in late years, they hire people with several year's working experience.

## 5.2 Agency:

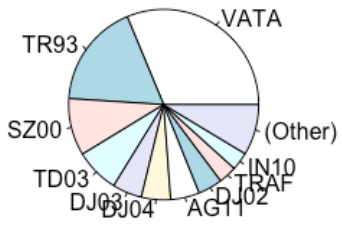


Figure2. 2001 Agency

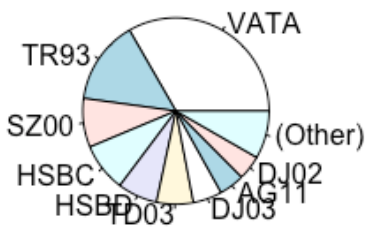


Figure3.2008 Agency

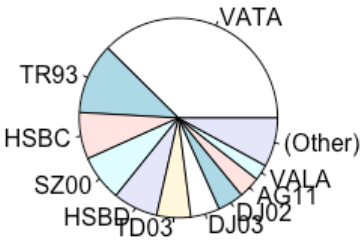


Figure4.2014 Agency

I use three pie chart to represent several agencies with most employees in different three years. It is

very clear to see the distribution of different agencies. In 2001, 2008 and 2014, first two agencies are VATA and TR93. In figure 2 and figure 3, the third largest part is SZ00 however it drop to fourth in figure 4. At the same time, HSBC and HSBD are not shown on figure 2, but becomes quite important in figure 3 and figure 4, especially figure 4 for HSBC.

By relating to the name of agency, we know that VATA--VETERANS HEALTH ADMINISTRATION, TR93--INTERNAL REVENUE SERVICE, HSBC-- TRANSPORTATION SECURITY ADMINISTRATION, SZ00--SOCIAL SECURITY ADMINISTRATION, HSBD--CUSTOMS AND BORDER PROTECTION. It seems like that USA government extremely cares about veterans' health problem and revenue service. For 14 years those two are still two largest part. But for VATA becomes bigger and TR93 becomes smaller. Maybe it's on account of two presidents' tax reduction policy. As for HSBC and HSBD, I think it is because government cares more about people's transportation security and consumer rights. People are most important thing all the time.

## 5.3 Education:

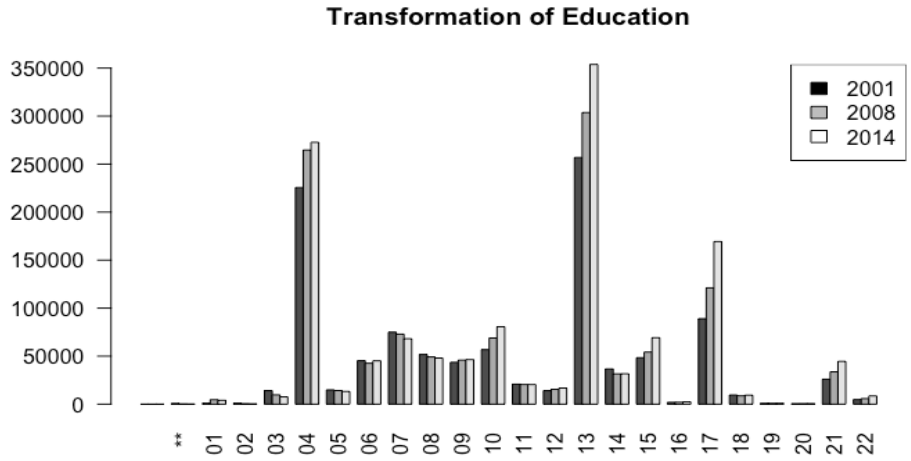


Figure 5. Barplot of Education

In figure 5, x axis means the education level, y axis represents employee numbers in each lecel. We can see that most people are level 4, 13 and 17. 4 represents HIGH SCHOOL GRADUATE OR CERTIFICATE OF EQUIVALENCY , 13 represents BACHELOR'S DEGREE, 17 represents MASTER'S DEGREE. As time goes by, people are willing to achieve higher education level than ever before working. Or more high-educated people are willing to work in government agency. It is very clear from the picture that 2014's amount are higher than 2001 and 2008's in group 13 or later.
Since education level is a group, I think barplot is good to compare values between years.

# 5.4 LOS:

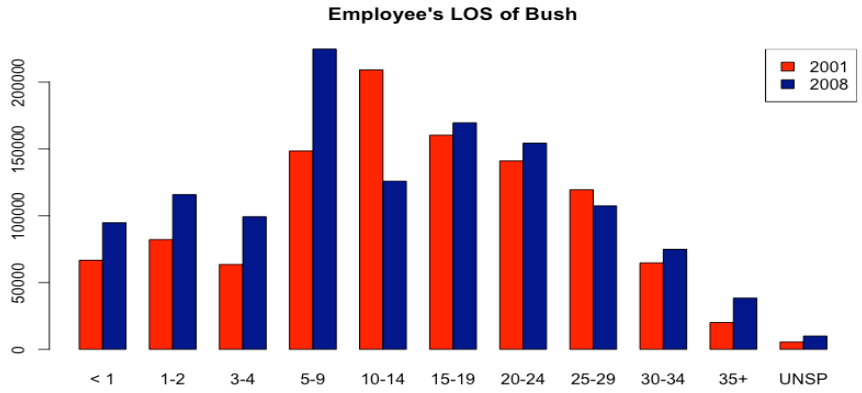


Figure 6. Barplot of Employee LOS in BUSH year

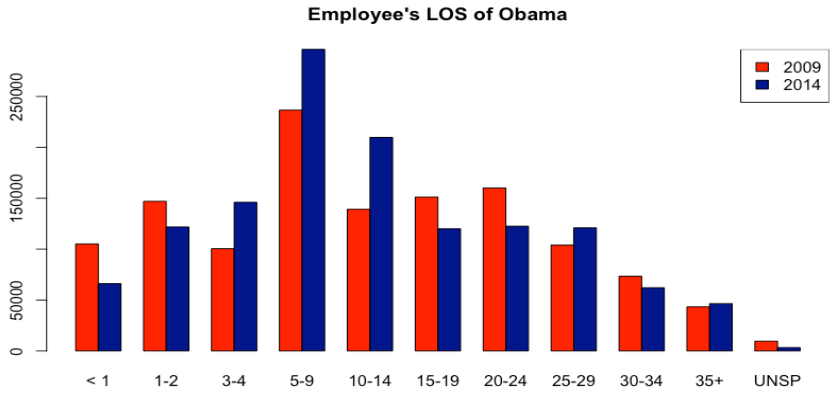


Figure 7. Barplot of Employee LOS in Obama year

In figure 6 and figure 7, x axis represents Length of Service year, y axis represents employee numbers of those groups. Figure 6 indIcates employee's LOS for first year and last year of president Bush. Figure 7 indicates Obama's. I would like to compare under two different government, if people were willing to work in government agency or not. But I cannot take other variables into consideration at a time, so other variables may also influence people changing job or retiring. I think this figure is meaning only at some aspect.

Comparing 2001 and 2008 in Bush year, almost each group increases employee number in 2008 except 10-14 and 25-29. And for Obama year 2009 and 2014, more groups have less employees in 2014. Concentration on group 10-14 for both Bush and Obama. It is completely converse. In Bush year, people in 2008 are less than 2001. In Obama year, people in 2014 are more than 2009. In my opinion, it is a bottleneck for everyone if staying in one company for 10 years without any promotion. So, maybe Obama government will promote people who work for 10 years. Similarly, maybe people in Bush time will be promoted if work for 20 years.

# 5.5 Pay:

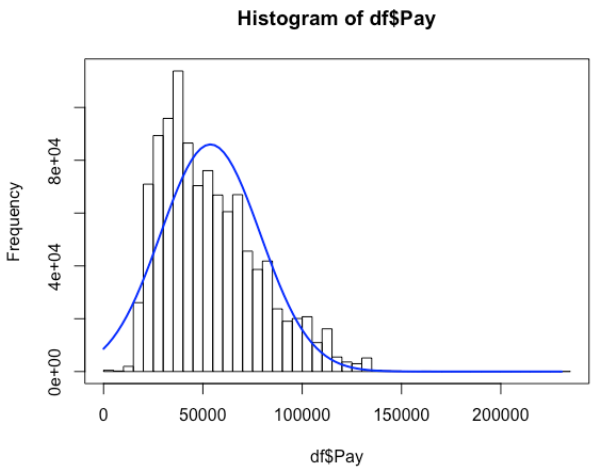


Figure 8. Histogram of 2001 pay

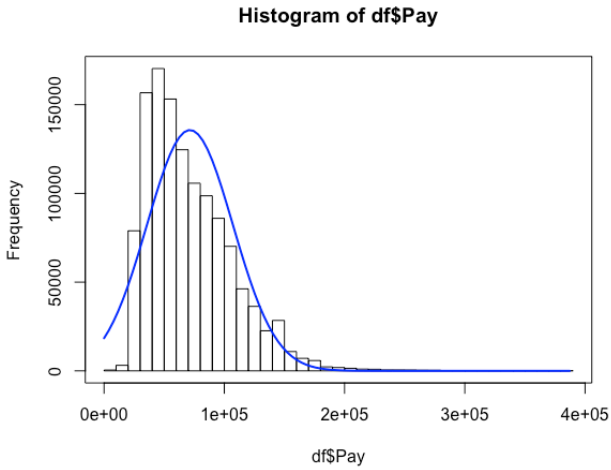


Figure 9. Histogram of 2008 pay

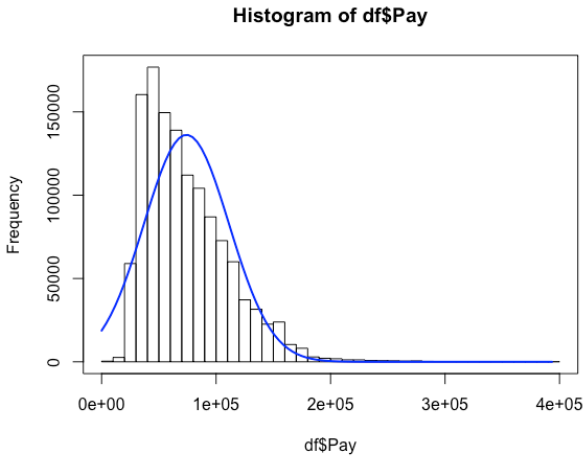


Figure 10. Histogram of 2009 pay

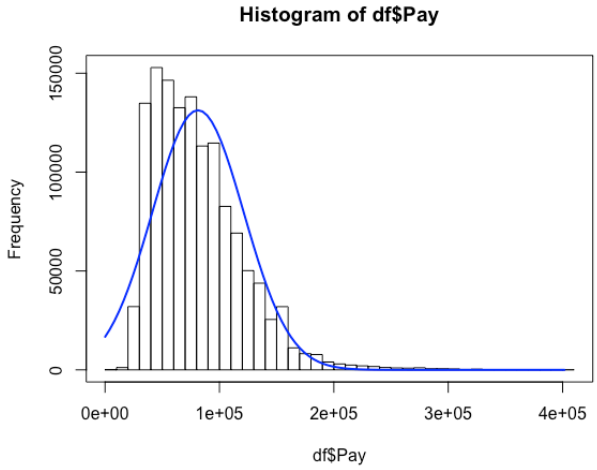


Figure 11. Histogram of 2014 pay

Those four figures' x axis means salary, y axis means number of employees. Figure8 and Figure9 focus on Bush year, figure 10 and figure 11 focus on Obama year. By comparing first two graph, it shows that people earn more in 2008. However, in figure 8, most people get pay at around $3500. In figure 9, most people get pay at around $40000. Also, histogram's curve match more with the normal distribution curve. Very rich people become less. Distribution of money become more concentrate.

Comparing figure 10 and figure 11, I find that money distribution become more concentration than 2009. Most people get salary amount range from 40000 to 10000. But very rich people become more than 2009.
In conclusion, 14 years pass over, people's salary become intensive, and match better with the normal distribution curve.

# 6 Relationship between variables

## 6.1 Education-LOS:

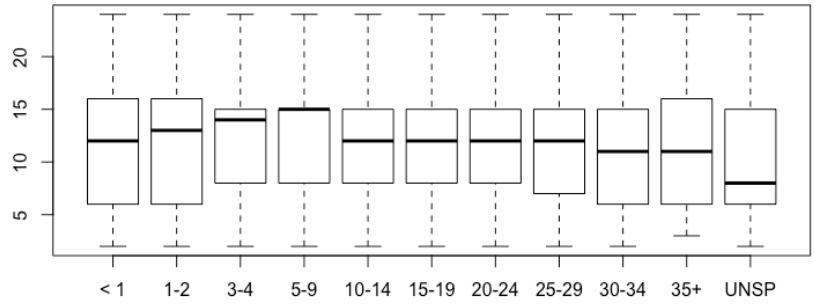


Figure 12. Boxplot of Education-LOS in 2001

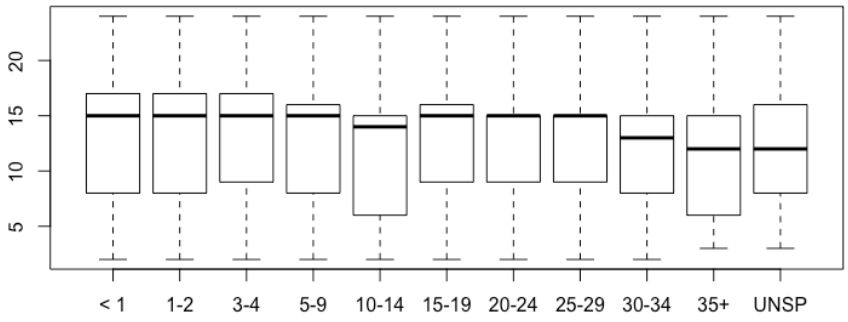


Figure 13. Boxplot of Education-LOS in 2014

These two figure expain relationships between Education and LOS. The x axis means education LOS, y axis means education level. From figure 12 and 13 we can see that almost every LOS group, the percentile of education level are similar. But in figure13, the median in each LOS group become higher than corresponding group in figure 12. Maybe because that after 14 years, people's education level increased a lot. The education level may not have strong effect on people's length of service.

## 6.2 Pay-Education:

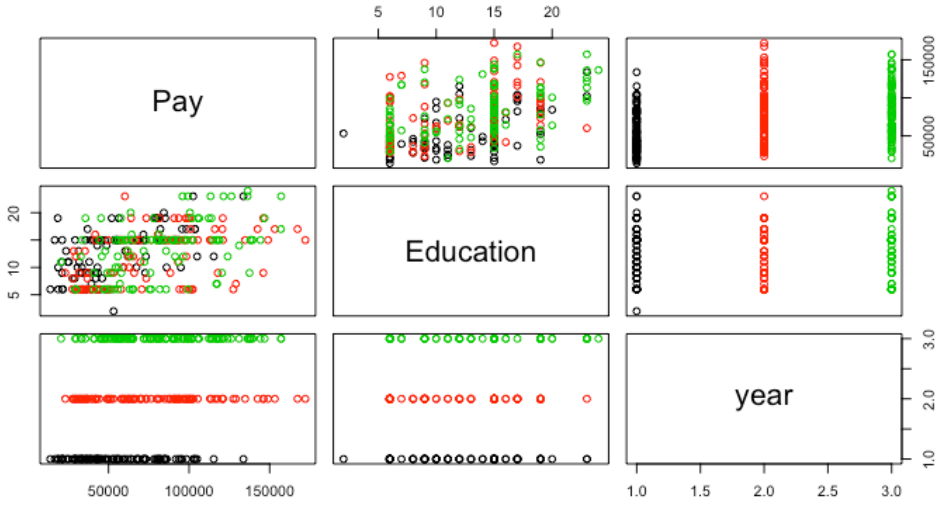


Figure 14.Plot of Pay-Education in three years

I choose 2001, 2008 and 2014 as my data because 2001 is the first year of president Bush. And 2014 is the last data year. 2008 is in the middle and it is also the last year of president Bush. So I sampled 100 data from each year, and plot figure to consider the relationship between pay and education. The black represents 2001. Red represents 2008, green represents 2014.

By looking at the picture located in row 2, column 1, we can see that most points focus on the left and bottm. It means most people are in low or medium education level and get low salary. But some red and green points disperse to right and top part in this picture. Comparing these red and green points, we can analyze that in 2008, extremely high education level is not required to get high salary. But in 2014, people getting high salsry having high education level.

Looking at top rifht figure, people get more salary in 2014 and 2008 compared to 2001. And money distribution is more intensive in 2014 than 2008.

The correlation between those two variables are shown below:

| | Pay | Education | year |
|---|---|---|---|
| Pay | 1.0000000 | 0.45686616 | 0.31887950 |
| Education | 0.4568662 | 1.00000000 | 0.01317149 |
| year | 0.3188795 | 0.01317149 | 1.00000000 |

Table 2. correlation table

From table2, we can see the education and pay are moderate related. Pay and year are moderate related, education and year are light related.

# 6.3 Pay-LOS:

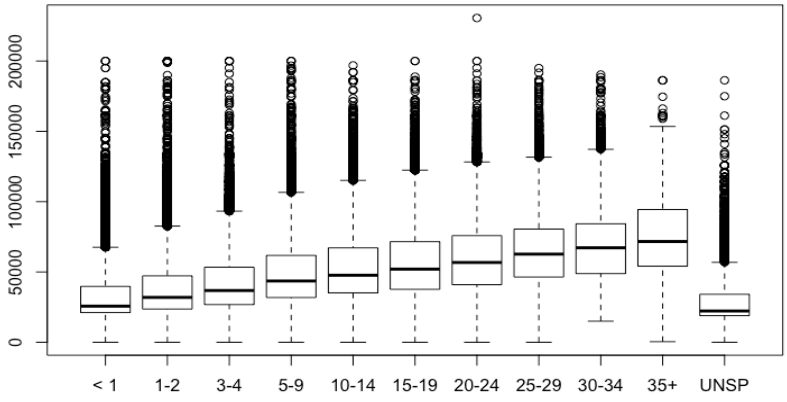


Figure 15. Boxplot of Pay-LOS in 2001

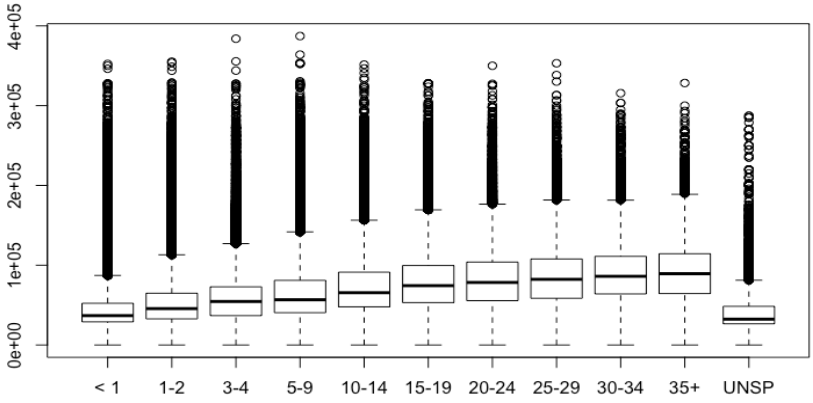


Figure 16. Boxplot of Pay-LOS in 2014

For these figure, x axis is LOS, y axis is pay. In both picture we can know that pay will be higher if LOS becomes larger when we dismiss some outliers, just focus on major part. These two has positive correlation. And also, people's salary become higher in 2014.

## 6.4 Age-Pay:

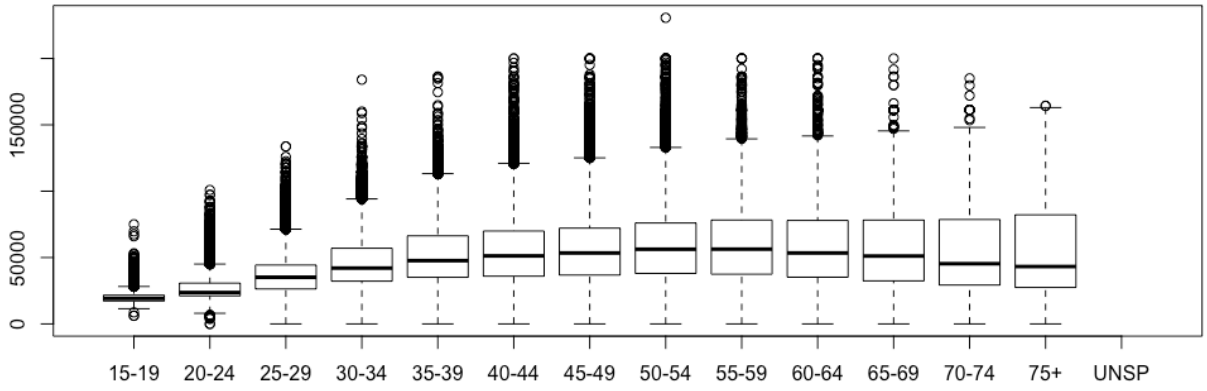


Figure 17. Boxplot of Age-Pay in 2001

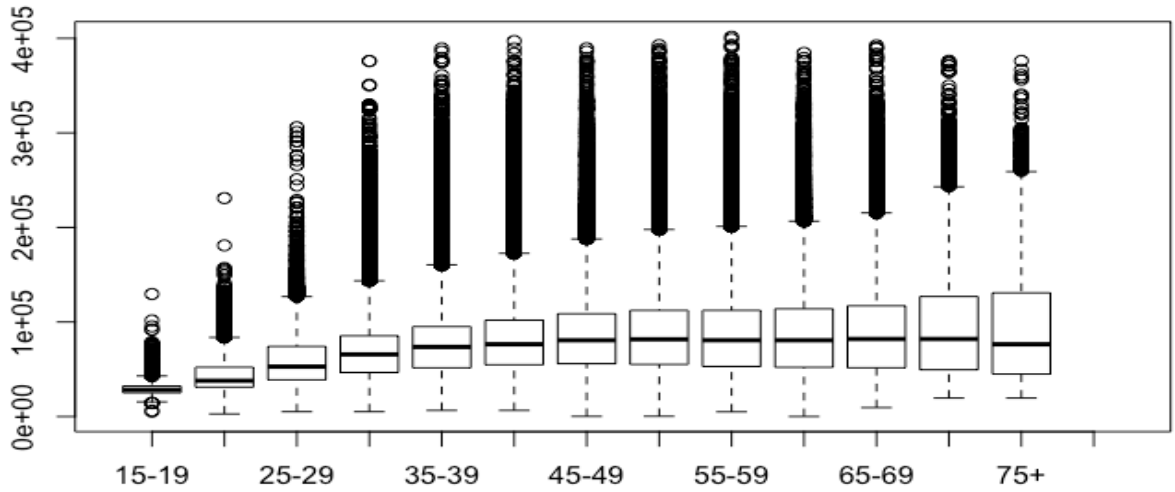


Figure 18. Boxplot of Age-Pay in 2014

The x axis is age group, y axis is pay. From both figure, it seems like group 50-54 has highest median of pay. It it not like the relationship between LOS-Pay, age and pay do not have positive correlation. Also in figure18, outliers becomes more compared to figure 17 when age more than 65. A factor of old people earn more than before, especially for 75+ people.

## 6.5 Agency-Pay:

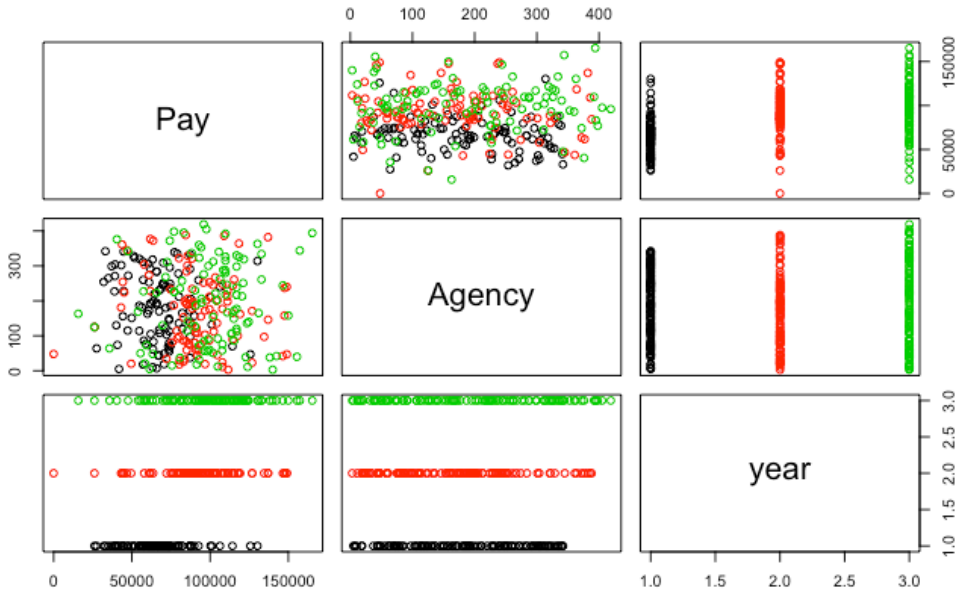


Figure19. Plot of Agency-Pay in three years

It represents agency's median salary hand out to employees. The black one is 2001, red one is 2008, green one is 2014. I sampled 100 agencies from each year, and plot them like above.

By looking at the picture located at row1, colomn 2, green points focus on the top, red points focus on the medium, back points focus on the bottom. It tells us agencies pay more to employees as time goes by.

The correlation between pay and year is

0.4872439

This is moderate positive correlation. So if time changes, agency's pay will follow change.

# 7.Data change over time

At first, I will set data into a data frame.

Then I will handle current data in a specific way that stored old data. If the variable is nominal, I can count it. If the variable is ratio/interval/order, I can calculate its mean/median

Then I will use cbind() function in R to record time of variable, use one column to record it. So I will not mess up with other data.

Next using rbind() function in R to combine these data with old data, and put them into one data frame. So it is convenient to do comparison with data in different generation. And data frame can be easily added or modified.

For over 40 years, the changes between different president are quite a lot. I think it is better to analyze dfferent periods of data, then doing comparing.

# 8. Conclusion

**Most important findings:**
In conclusion, analyzing nominal variable, counting is a good idea. As for interval/order variable, boxplot is a good method. For ratio variable, histogram is good for use.
Agencies pay more as time grows.
Pay has positive correlation with LOS.
Education and pay have positive correlation with each other.
Age between 50-54 earn most salary. Old person become earn more than before.
People's education level become higher.
In all group of LOS, their 1st and 3rd percentile of education level are similar.

Three advantages:
1. I think my report is very beautiful, including type of words and arrangement of pictures.
2. I try my best to find relations between variables. And find changes on a variable in different years carefully.
3. Analyzing objectively

Three disadvantages:
1. Incoherent language. The expression is not good.
2. Not to pay attention to many other variables, it makes me miss some interesting thing
3. A little bit verbose, not very explicit and straight to describe one thing.

# 9.References

http://michael.hahsler.net/SMU/EMIS7331/data/federal_employment/project1.html
http://us-presidents.insidegov.com/compare/2-39/Barack-Obama-vs-George-W-Bush
https://en.wikipedia.org/wiki/September_11_attacks
https://www.thebalance.com/2008-financial-crisis-3305679
https://github.com/jakecarlson1/data-mining-projects/blob/master/project-1/project1.R