

**ASSIGNMENT FRONT SHEET**

**Course Name: ALY6010 71904 Prob Theory and Intro Stats**

**Professor Name: Dr. Morteza Ziyadi**

**Student Name: Dong Quoc Tuong (Lukas)**

**Student Class: Fall 2019 CPS Term: A. 2019**

|  |
| --- |
| **Module 1 Project: Basic Univariate Statistics, Graphical Methods, and Communication of Data**  **Completion Date:September 24th Word Count: 800 Due Time:12:00am** |

**Statement of Authorship**

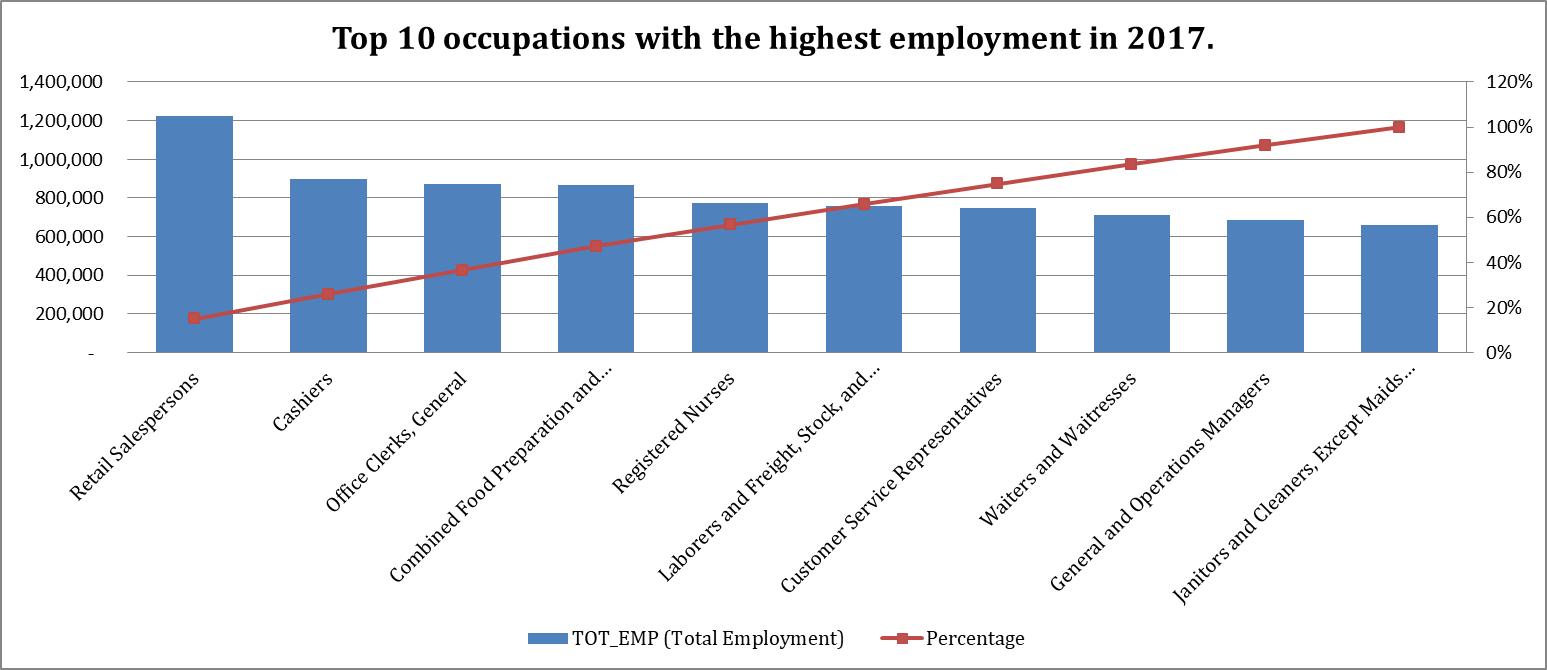
*I confirm that this work is my own. Additionally, I confirm that no part of this coursework, except where clearly quoted and referenced, has been copied from material belonging to any other person e.g. from a book, handout, another student. I am aware that it is a breach of Northeastern University’s regulations to copy the work of another without clear acknowledgement and that attempting to do so renders me liable to disciplinary procedures. To this effect, I have uploaded my work onto Turnitin and have ensured that I have made any relevant corrections to my work prior to submission.*

**Tick here** to confirm that your paper version is identical to the version submitted through Turnitin

In this paper, we will take a look at the results to understand the employment environment in the US in 2017.

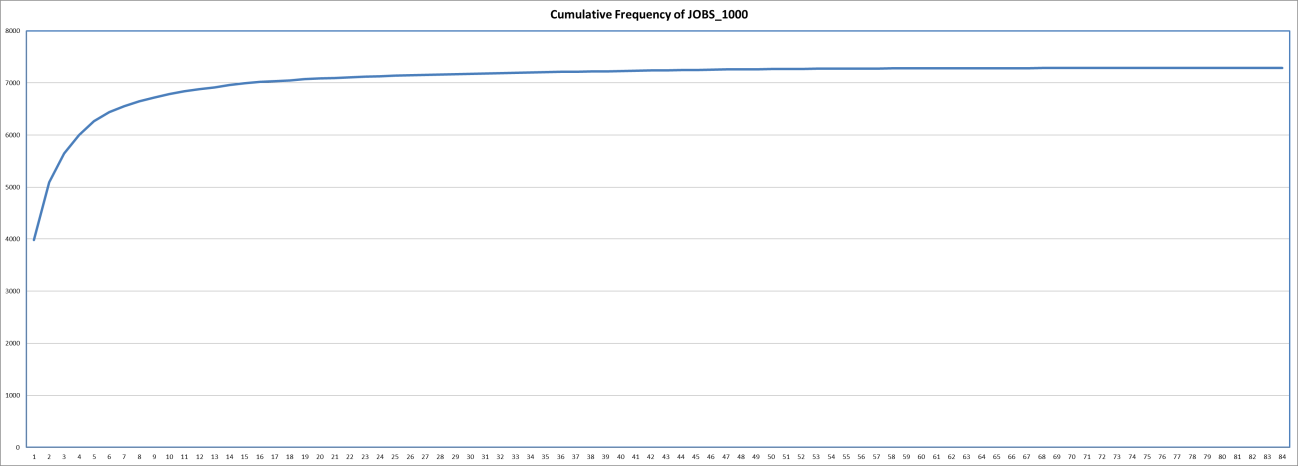
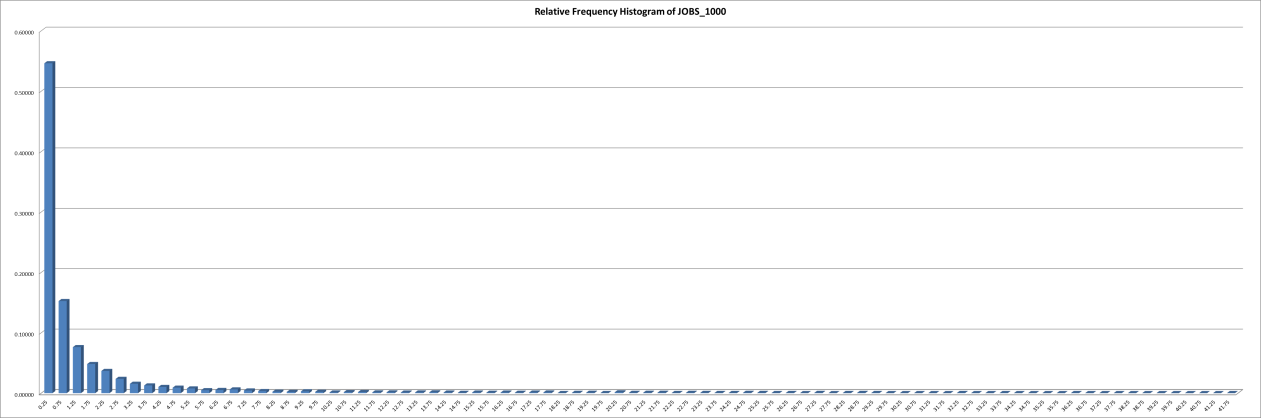
**Question 1:**

Disparity in term of number of working people between the most popular professions in America remained insignificant. Retail Sales was the most in-demand employment with nearly 1.3 million workers, twice the amounts of people in sanitary sector such as Janitors or Cleaners.

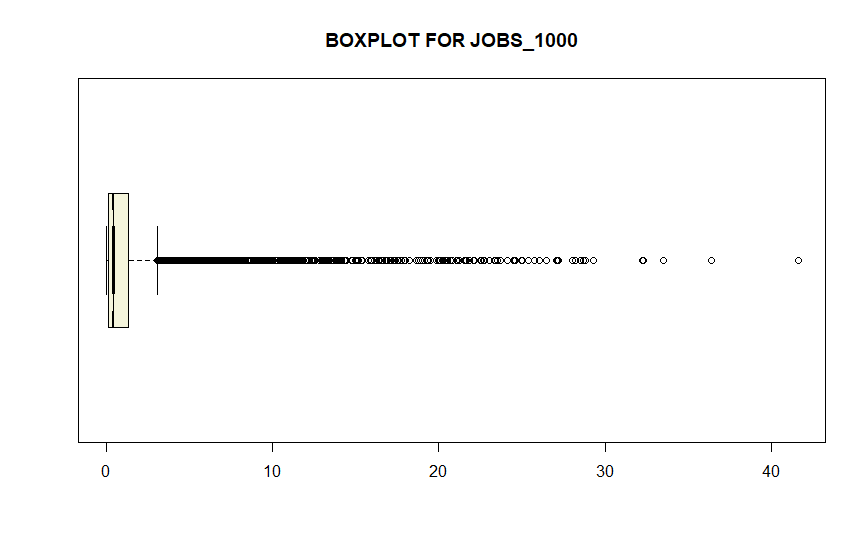


**Question 2,3 :**

These 2 graphs below demonstrated the number of employments in the given profession per 1000 jobs in all areas across America. The higher the “JOBS 1000” number is, the higher chance that particular area is dedicated for a specific industry or profession. The relative frequency distribution histogram from the “JOBS\_1000” is positively skewed. There are extremely few data plots that can be found on the right of the graph, creating a very long “tail” for the graph. The mean (“average” value , 1.47) is also pulled towards the tail and greater than the median (“middle” value ,0.41). The same can be said about Cumulative Frequency distribution chart where it stayed unchanged after dramatic increase in the beginning. Due to such graph’s unique look, it is safe to say there is almost no job that is excessively sought after over another in one area. Retail Sales is the exception since job markets in big cities like Miami, New York, Detroit, Dallas, Seattle, San Francisco, and Chicago are dominated by this particular profession, Higher population in urban areas equals higher purchasing power. Thus demand for retail specialist flourishes. Personal Care Aiders are surprisingly popular in Los Angeles and San Francisco. A potential answer to such phenomenon is California’s current homecare crisis, driven by the aging of the baby boom generation and favor of the affordable homecare over costly nursing homes’ service(Thomason & Bernhardt, 2017)

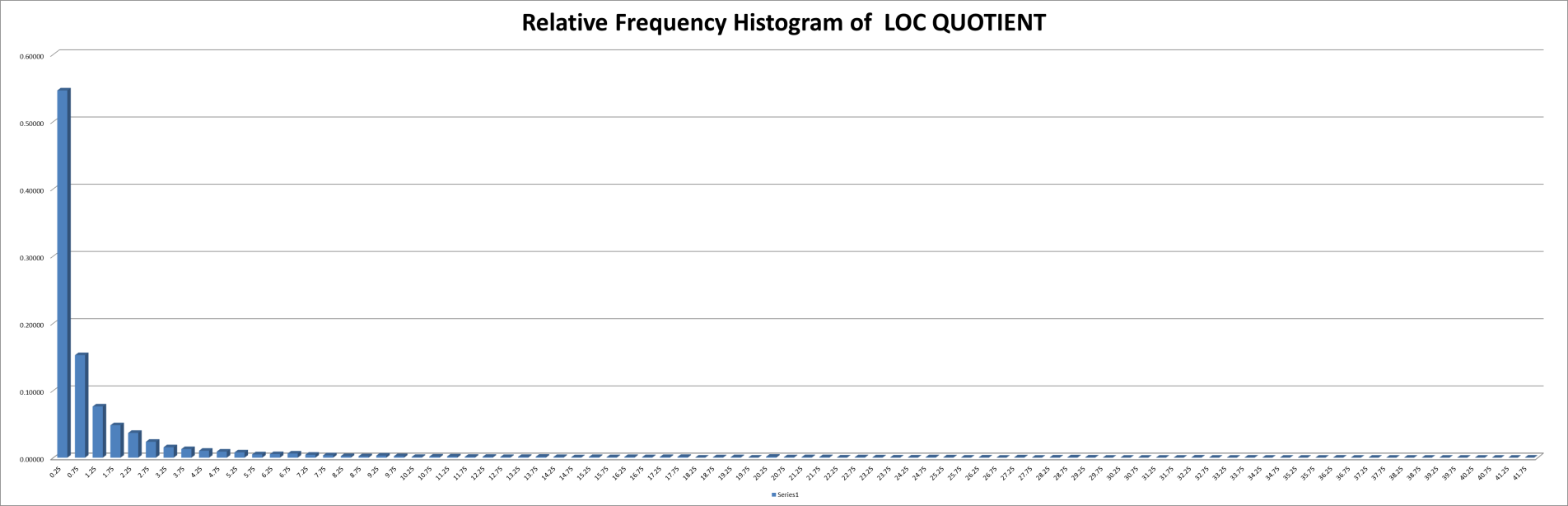


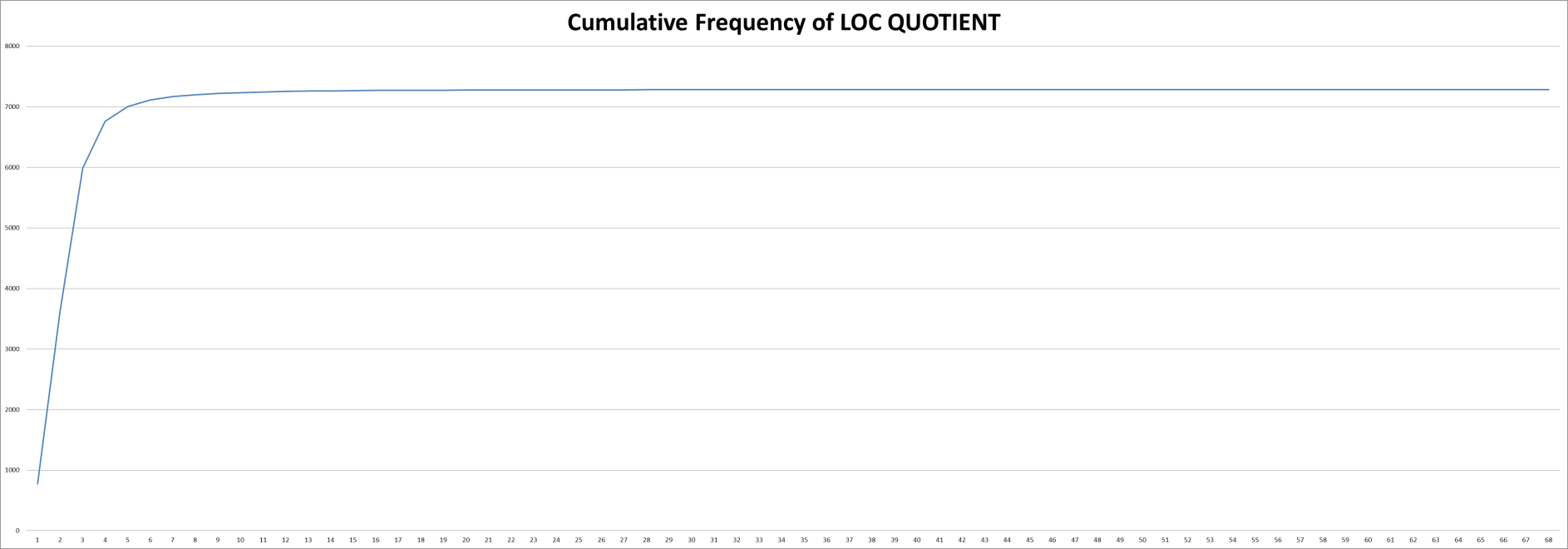
**Question 4,5**

Perform numerical descriptive statistics

From a visual point of view, there seems to be a lot of outliners on the Boxplot graph. But in reality, there are only 827 outliners; just one out of ten variables is outliners. All the outliners lie on the right of the graph with an abnormal distance from other values. The further we moved away from the Upper bound, the more scattered these extreme outliners become. Given to the sizable amount of outliners, some of which equal to each other, it would be advisable to count all the outliners as part of the study rather than discard them. (Rousseeuw & Hubert, 2011). The only exception could be the five outliners that really distance themselves from the rest.

**Question 6**

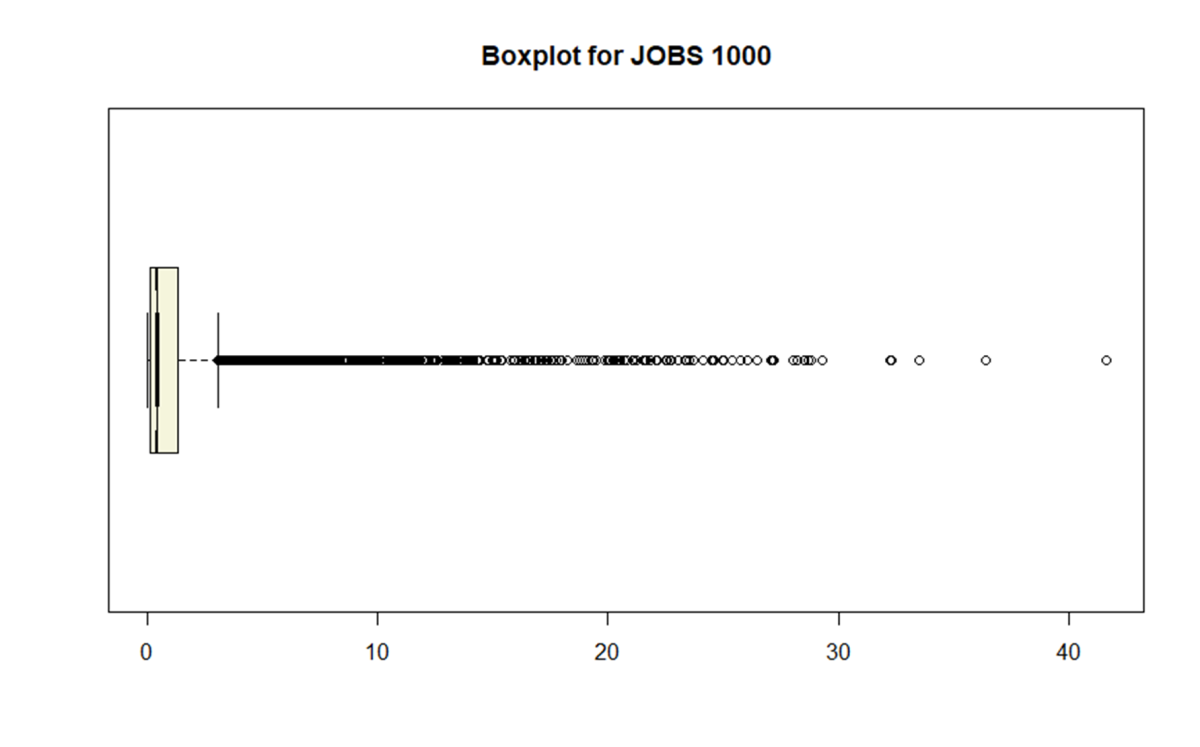


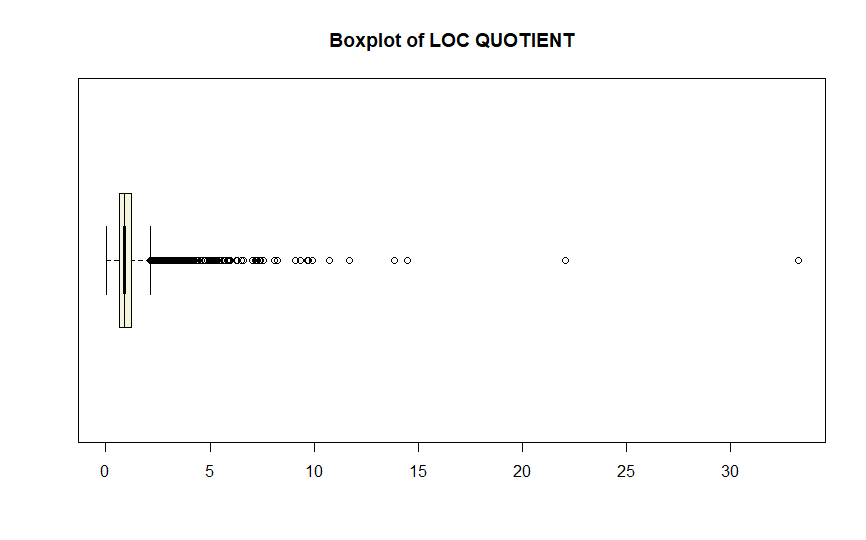


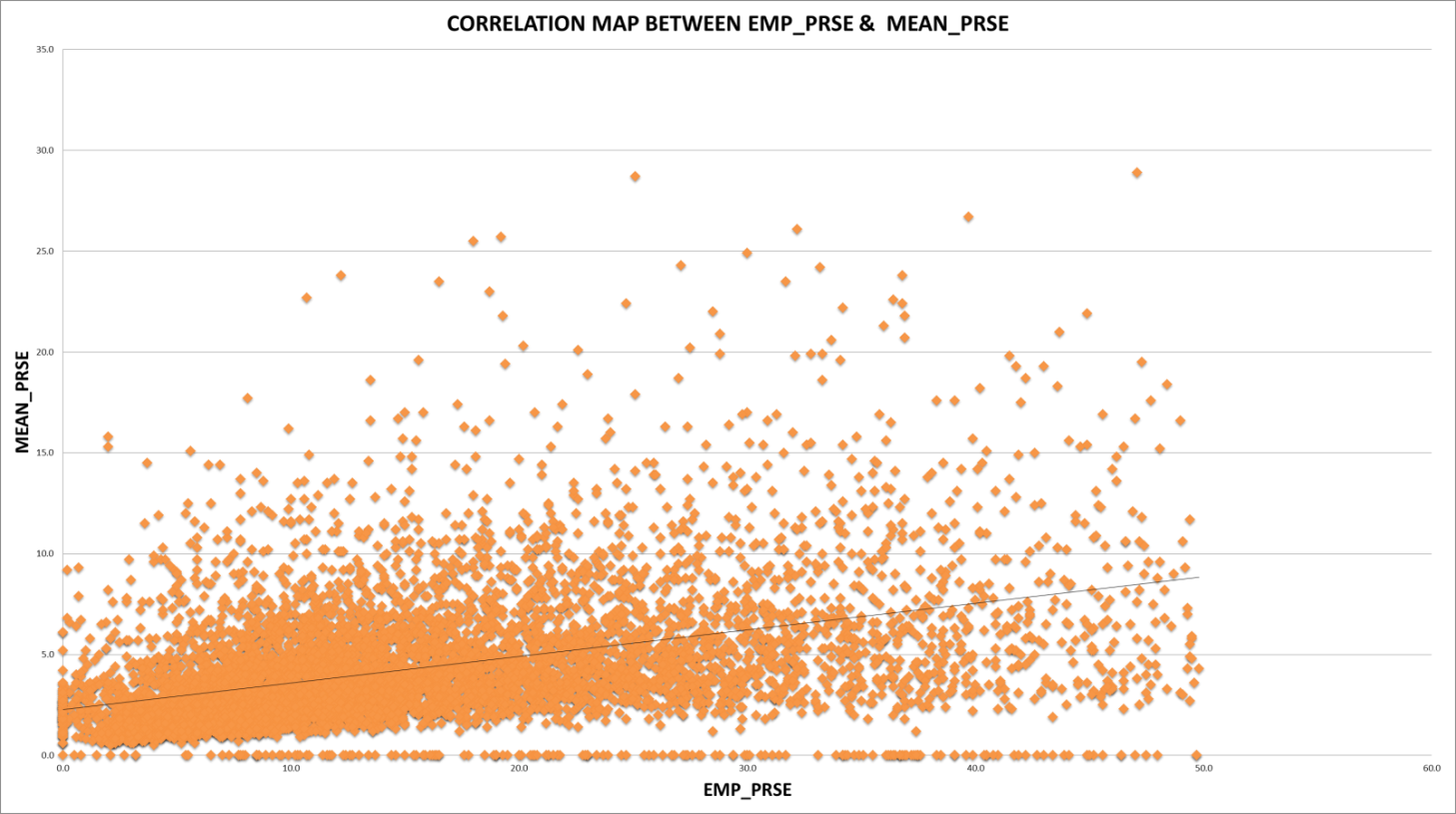
These 2 graphs above of the LOC QUOTIENT figures are nearly identical to those of JOB 1000. The relative frequency distribution histogram from the “LOC QUOTIENT” skewed dramatically to the right. This is very surprising because ”LOC QUOTIENT” illustrates the ratio of the job’s share of employment of the given area and the whole country. It is easy to say that popularity of a job in a specified area correlates strongly with its popularity at the national level, but it is not correct. Because Economists and Political Scientists in Washington have the highest ratios.

The difference between two sets of data becomes more evident once we have a look at the measurements. “LOC QUIENT” has larger Minimum, Q1,Q2, L Bound but smaller Maximum,Q3, U Bound compared to those of “JOBS1000”. Consequently, the Box Plot of “LOC QUIENT” is smaller but much evenly distributed compared while JOBS1000’s is bigger than size yet extremely right skewed. ”LOC QUOTIENT” dataset‘s number of outliners is less than half of that of the “JOB 1000”. Additionally, the outliners of “LOC QUOTIENT” do not stay too far away from the quartier range with the only exceptions.







**Question 7** The map above illustrates the correlation between the precision of statistics of the mean wage and employment of an occupation in a particular area. The variables gather at the bottom left of the graph means that the majority of both studies are highly accurate. The linear regression goes upward in the East North direction indicates that there is noticeable correlation between the majorities of variables of these two numbers. Nonetheless, the further the line goes upward, the slimmer it gets, suggesting that the positive correlation will be decreased once the relative standard error of these two axis increase dramatically or if they are outliners. Furthermore, the accuracy of mean wage statistic is extremely high even though their “Employment” statistics fluctuate greatly, evidently shown by the number of markers at bottom of the chart. The same cannot be said vice versa with only a few markers stay on the left axis. In conclusion, statistic of the mean wage is much more reliable compared to the statistic of employment, which means that the precision of the dataset of JOBS 1000 and LOCQUIENT are questionable

**References**

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Minining and Knowledge Discovery*, (January). https://doi.org/https://doi.org/10.1002/widm.2

Thomason, B. S., & Bernhardt, A. (2017). *California’s homecare crisis : raising wages is key to the solution*. *UC Berkeley Center for Labor Research and Education*. Berkeley. Retrieved from UC Berkeley Center for Labor Research and Education