MSDS 7330-403: SQL Project

Laura Bishop and Matthew Przybyla

# Introduction

This is a Standardized Query Language (SQL) project for Masters Science in Data Science (MSDS) class 7330-403. The dataset chosen is White House Salary Data located at <https://opendata.socrata.com/Government/2010-Report-to-Congress-on-White-House-Staff/vedg-c5sb> .

The Baseball salary example in MSDS 6371 – Foundational Statistics inspired the choice of this data set. The White House Salary data set is the basis of a report presented to the United States (US) Congress detailing salaries of employees working in the Obama administration White House in 2010. While 2010 is not the only year White House salary data is collected, and presented to the US Congress, it is the data set year chosen for this project.

The ‘after presentation’ conversation in class about the White House and government positions, prompted this brief explanation of “serving at the pleasure of the President”. According to Article II, section 2, of the US Constitution gives the President the power to nominate public officials with the Advice and Consent of the Senate.”[[1]](#footnote-1) The most notable confirmations that require consent are of Cabinet secretaries and Supreme Court Justices. The President has the power “to unilaterally appoint over 350 people to high-level positions within the federal government.”[[2]](#footnote-2) The Government Accountability Office indicates that most presidential appointee salaries are between $99,000 and $180,000 a year with full benefits.[[3]](#footnote-3) This makes the data interesting to understand how much White House employees make.

# Data Explanation

The environment for analysis is MYSQL on IBM Bluemix. There are 469 rows and five variables in the White House Salary data set for 2010, including the following:

* employee name,
* employee status,
* salary,
* pay base,
* position.

For the most part, this data is clean with no significant missing values and is in Microsoft Excel ‘.csv’ format. It is important to note that a job category or General Schedule (GS) job ranking did not exist in this data set. The ‘pay base’ variable only had the value ‘per annum’. This drove the team to explore the data and determine which position titles appeared to carry the most significance. From this analysis, the following titles were chosen for the initial review:

1. Legal Assistant 7. Director
2. Deputy Assistant 8. Analyst
3. Assistant Director 9. Chief of Staff
4. Deputy Director 10. Assistant to the President
5. Deputy Associate 11. Deputy Assistant to the President
6. Policy Assistant

# Data Analysis & Results

Data analysis began with importing the data and normalizing it into two tables: Employee, which contains EmployeeName, EmployeeStatus, and PositionTitle; and SalaryTable, which contains Salary and PayBasis. The two tables were joined by a foreign key in the Employee table, which contained the SalaryTable primary key id.

Exploratory analysis focused on PositionTitle and related salary. For example, the maximum, minimum and average are listed below. Note the minimum salary is $0, meaning at least three people were “volunteering” at the pleasure of the president.

* Maximum Pay $179,700
* Minimum Pay $ 0
* Average Pay $ 82,961

The title that earned the least was Legislative Assistant at salaries ranging from $42,000 to $45,000. In 2010, a GS level 7 job earned from $33,979 to $44,176 and a GS level 8 $37,631 to $48,917.[[4]](#footnote-4) Clearly the Legislative Assistants are in the GS 7 or 8 category.

The histogram below illustrates that most employees in the White House earn approximately $45,000.



Figure 0‑1 Histogra2010 Salary

In contrast, those with “Deputy” in their title earned at the top of the pay scale ($102,000 - $79,900) or about 3 times as much as those on the lower ends of the pay scale.

The top GS level pay grade in 2010 is 15, earning $99,628 to $129,517.[[5]](#footnote-5) It is inferred that those above the GS level 15 maximum pay are working at the “pleasure of the President”.

The Kaggle analysis of the White House Salary Data in 2016 and 2017 show very similarly shaped curves. Much of the similarity is due to those employees who are GS level ranked and are paid according to that pay structure each year. Thus, the expectation is not to have remarkable differences in salary ranges each year.

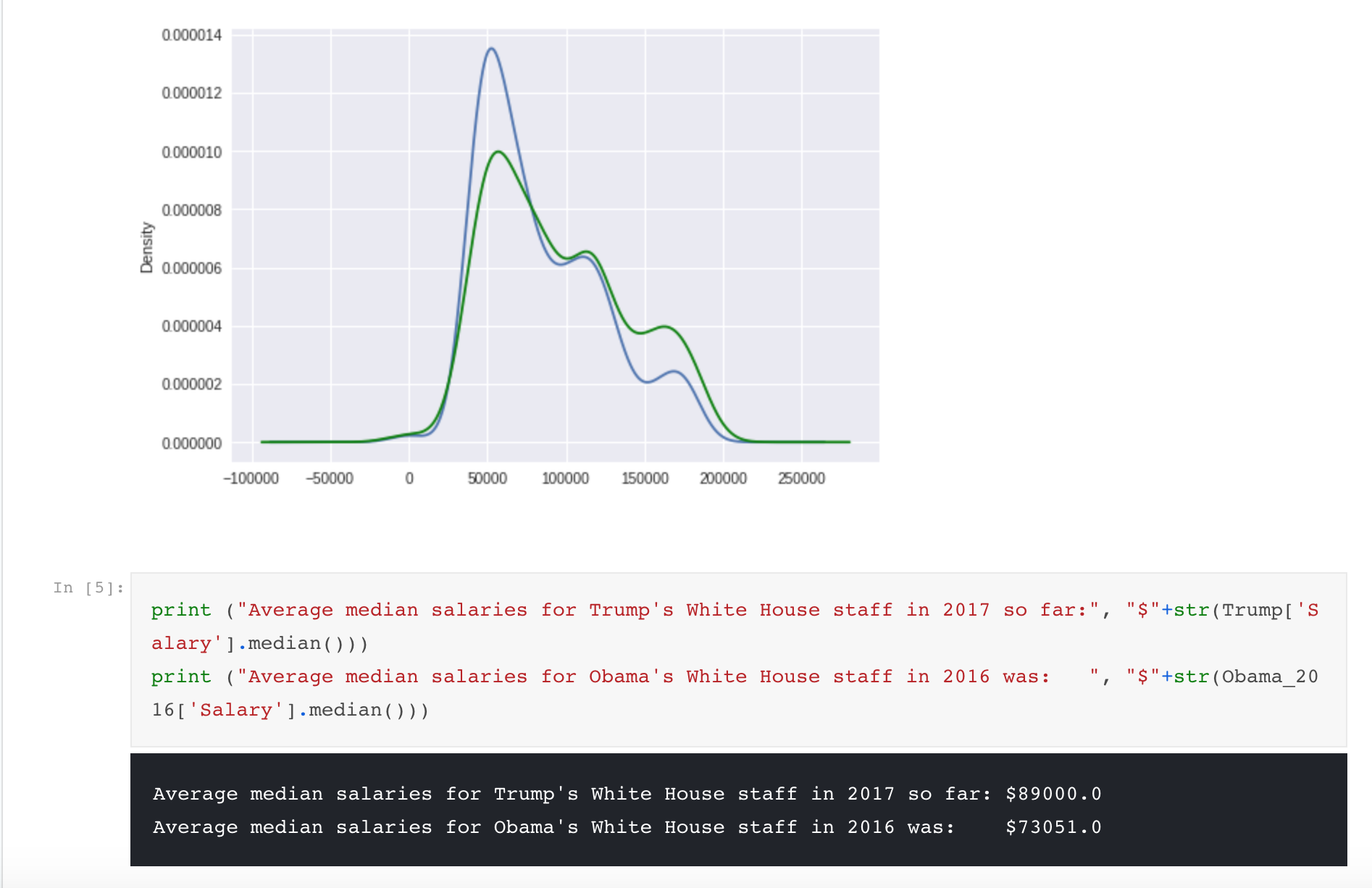


Figure ‑ Average Salaries for 2016 and 2017

# Conclusion

From a SQL learning perspective, this project was eye opening in how to import and segment off data, create tables and perform initial analysis of the data. This effort is harder than it looks and highlights the expertise and professionalism of the data base specialist in an organization.

From a data perspective, this analysis highlights the range of salary positions in the White House, which is similar to many large organizations. The White House salary range happens to be right skewed versus normally distributed in its employee salaries.

# Code

<https://github.com/AustinLVB/7330SQLclass/tree/master/Final%20Submission>

1. “Serving at the Pleasure of the President”, <https://www.archives.gov/publications/prologue/2005/winter/senate-nominations.html>

   and General Records of the United States Government, RG11 [↑](#footnote-ref-1)
2. <https://www.thoughtco.com/presidential-appointments-no-senate-required-3322124> [↑](#footnote-ref-2)
3. Ibid. [↑](#footnote-ref-3)
4. <https://www.federalpay.org/gs/2017> [↑](#footnote-ref-4)
5. Ibid. [↑](#footnote-ref-5)