Data Professional Salary Survey

Jason Tompkins

1/18/2021

The data are the result of three surveys completed in 2017, 2018, and 2019. The original source of the data was brentozar.com . Brent Ozar is a Database Administrator and online content creator. The data set contains 6893 observations of 29 variable. One thing to note about survey data is that the responses carry the biases of the respondents. Rather than approaching this data as information that was gathered by a uniform process, it is more precise to approach it as the information that respondents want to convey. In this context it is possible that respondents reported their salaries as slightly higher or responded humorously to a question that they found unimportant. For example one survey taker reported their gender as an “Apache helicopter”. This doesn’t invalidate the fact that the majority of respondents identified as male.

The analysis below seeks to answer the question “Can this data be used to gain insight into attributes that result in a higher than average salary for Data Professionals”.

Exploratory Data Analysis

Descriptive statistics were calculated (shown below). It became obvious early on that some extreme outliers were making it difficult to analyze the data, and in some cases impossible to visualize. The salaries above $500,000 and below $1000 were removed to better facilitate analysis. The outliers were inspected, but the attributes were consistent with observations with salaries closer to the average. It’s likely that the qualities of these individuals that demand such a high salary were not captured by the survey.

Before removing outliers	After removing outliers
The average salary is $92,458.66 The std dev of salaries is $55,765.26 The median salary is $90,000.00 The maximum salary is $1,450,000 The minimum is 0 The middle 50% of salaries fall between $65,000 to $115,000	The average salary is $91,151.88 The std dev of salaries is $41,425.04 The median salary is $90,000.00 The maximum salary is $500,000 The minimum is $1050 The middle 50% of salaries fall between $65,000 to $115,000

A few visualizations were generated to better understand the data. The majority of respondents were full time employees, had a bachelor’s degree, and work in the private sector. Next, a histogram was created to visualize the distribution of salaries. The distribution is skewed towards the average salary of $91,151.88 with a long tail toward the maximum salary of $500,000.

A box plot was produced to create a comparison between the salaries of data professionals that worked with different database platforms. The majority respondents work with Microsoft SQL Server, but the respondents who work with databases like MongoDB and Elasticsearch seem to have higher average salaries.

Finally, a scatter plot shows the positive relationship between salary and the years of experience respondents had in their current job.

Linear Regression Model

After the exploratory data analysis, a linear model was created to predict salary based on certain attributes. The attributes chosen as independent variables were; manages staff, country, primary database, years with this type of job, employment status, job title, other people on your team, database servers, education, education is computer related, certifications, and employment sector. The model accounts for 52.2% of the effects on salary, according to the adjusted R-squared value. The formula that was generated is below.

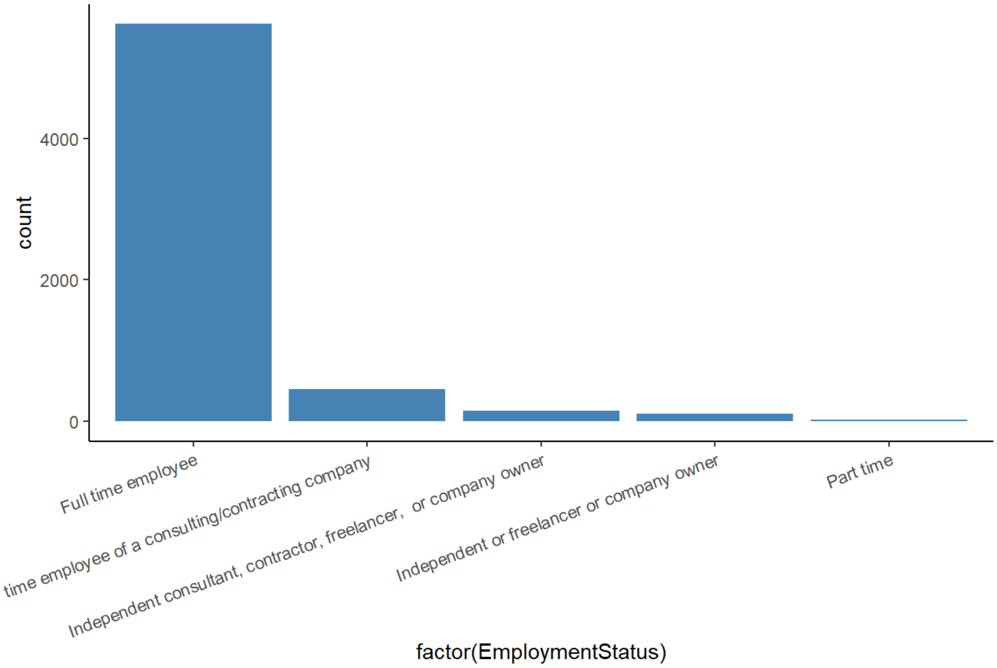
Predicted Salary = -23,224.07 + Country + Employment Status + Job Title + Education + Certification + Sector + (Number of years working that type of job * 1007.5)

In an effort to interpret the linear model a scenario was created. This scenario used some attributes that were less frequent in order to find a higher-than-average salary. The specific attributes are listed below.

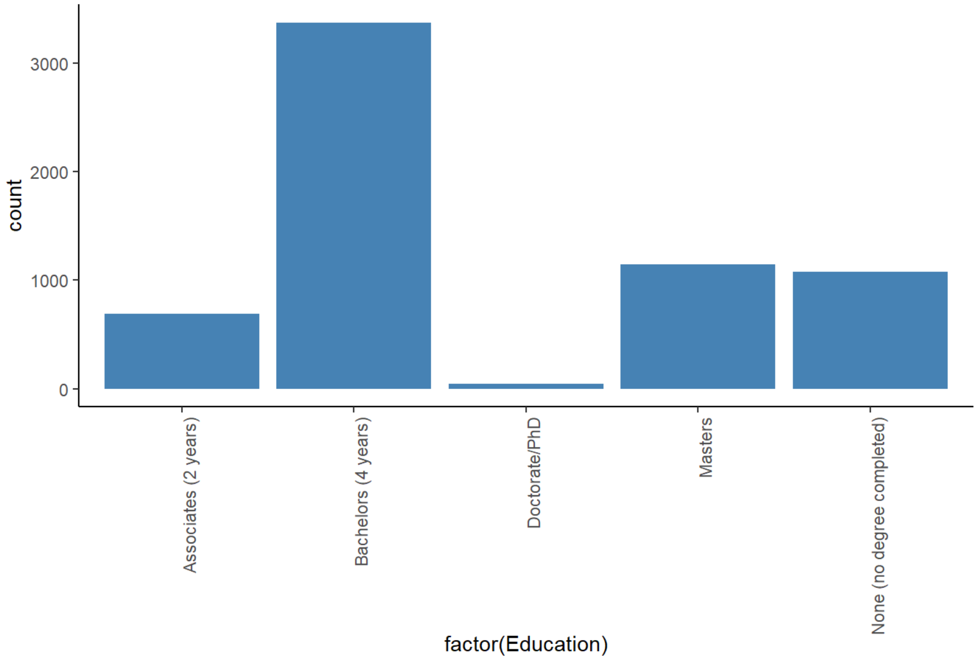
- Attributes
    - Country = US
    - Employment Status = Ind Contractor or Company Owner
    - Job Title = Data Scientist
    - Education = Master Degree
    - Certification = Yes but expired
    - Sector = Federal Gov’t
    - Number of years experience = 3

The resulting predicted salary was $193,905.

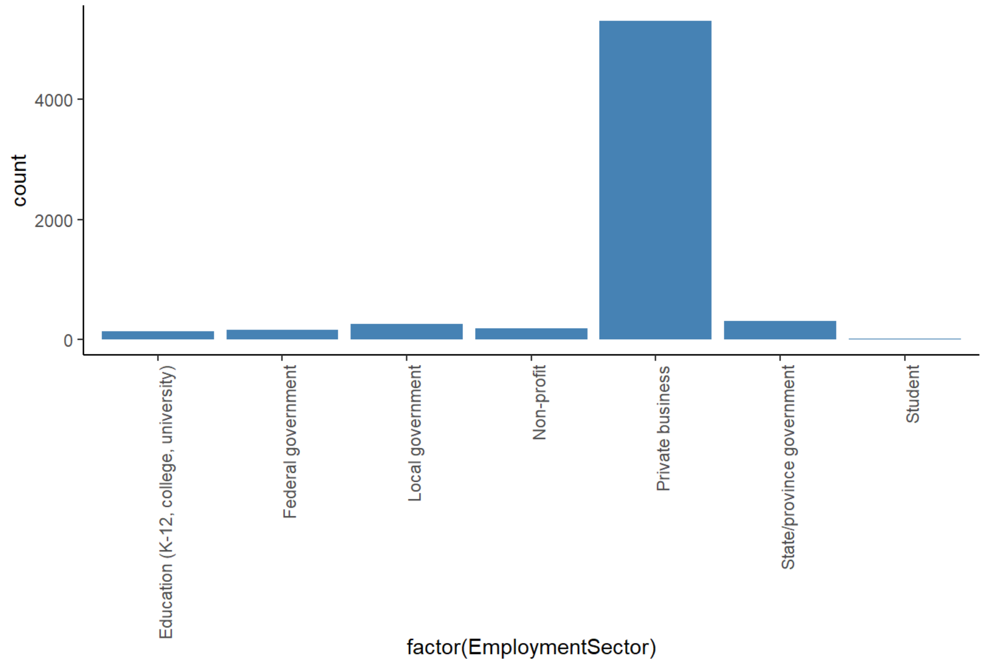
EDA Results

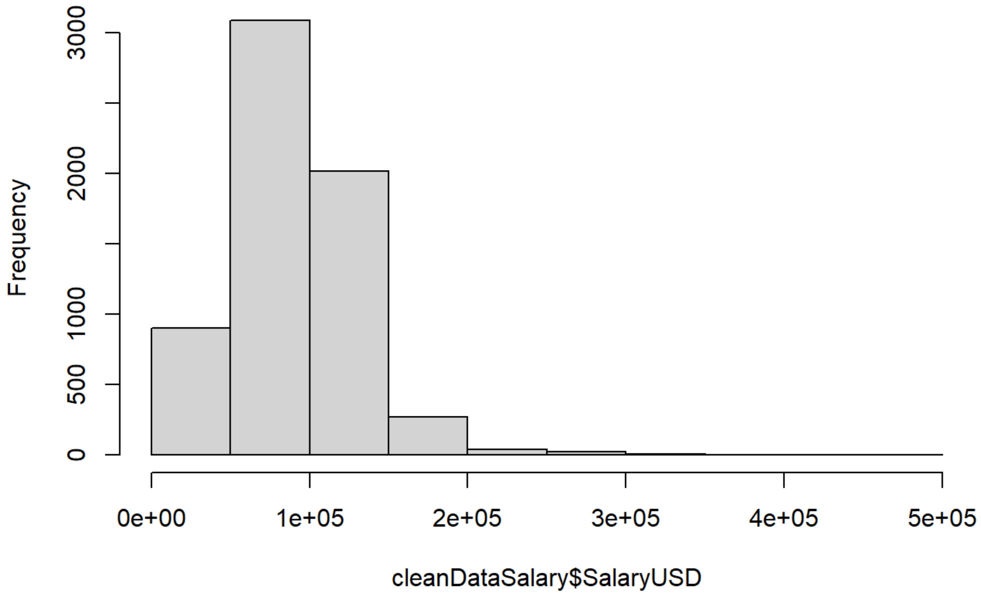


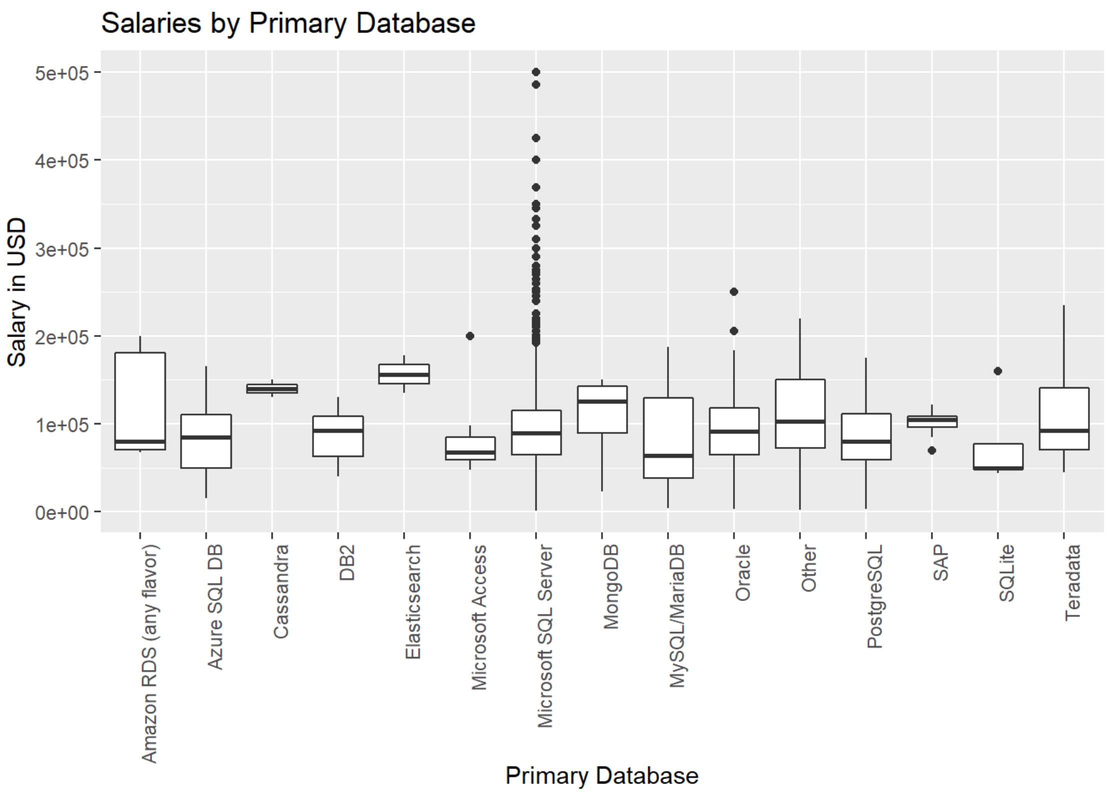
Employment Status

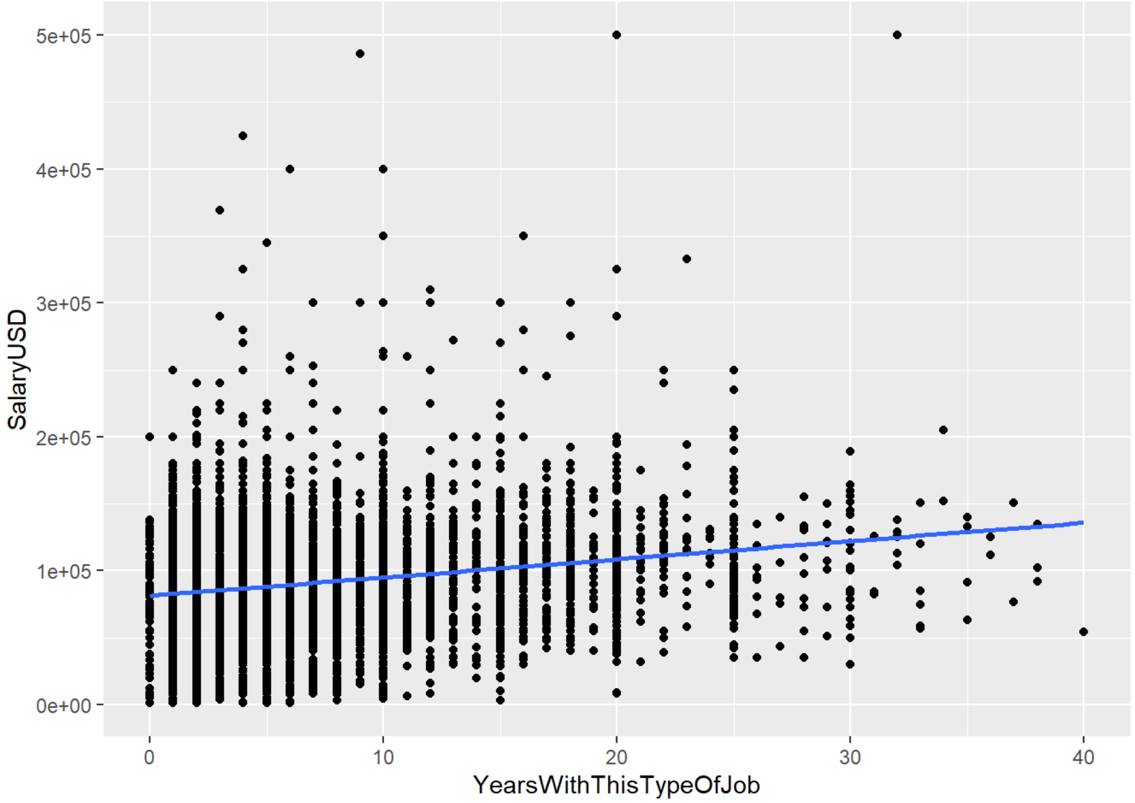


Education


Employment Sector


Histogram of cleanDataSalary$SalaryUSD


Salaries by Primary Database
Salary in USD
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
Amazon RDS (any flavor)
Azure SQL DB
Cassandra
DB2
Elasticsearch
Microsoft Access
Microsoft SQL Server
MongoDB
MySQL/MariaDB
Oracle
Other
PostgreSQL
SAP
SQLite
Teradata
Primary Database


SalaryUSD
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
10
20
30
40
YearsWithThisTypeOfJob

Linear Model Results

```
## Residual standard error: 28630 on 6198 degrees of freedom
## Multiple R-squared:  0.5331, Adjusted R-squared:  0.5223
## F-statistic: 49.15 on 144 and 6198 DF,  p-value: < 2.2e-16
```