Story_4

Alex Khaykin

2023-10-19

INTRODUCTION

According to the US Bureau of Labor Statistics, the median pay for data scientists was $103,500 per year or $49.76 per hour. The projected growth for this career and related fields is considered to be one of the highest in the US, but what do data scientists and other data practitioners(data analysts, business analysts, data engineers) actually make?

One source often used to gauge salaries are job listing sites such as glassdoor.com, however, these are often posted as salary ranges and may not accurately reflect the salaries drawn by such employees in the field. Instead, I will use data from Ask a Manager blog which polls actual employees across a variety of fields, and may more accurately represent the true salaries of data practitioners. This survey was opened in April of 2023, thus I can compare it to the 2022 median salary from the US Bureau of Labor Statistics.

Accessing the data The survey data are openly available via a Google Sheet, and includes information on industry, job title, salary, geographic location, on-site vs. remote work options, years of experience, highest level of education attained, gender and race.

```
dat <- read_sheet("https://docs.google.com/spreadsheets/d/1ioUjhnz6ywSpEbARI-G3RoPyO0NRBqrJnWf-7C_eirs/
```

```
## ! Using an auto-discovered, cached token.

##   To suppress this message, modify your code or options to clearly consent to
##   the use of a cached token.

##   See gargle's "Non-interactive auth" vignette for more details:

##   <https://gargle.r-lib.org/articles/non-interactive-auth.html>

## i The googlesheets4 package is using a cached token for 'akhaykin81@gmail.com'.

## v Reading from "Ask A Manager Salary Survey 2023 (Responses)".

## v Range 'Form Responses 1'.
```

```
nrow(dat)
```

```
## [1] 17034
```

There are 17,033 total responses to the survey which I will filter for data practitioners and only those paid in USD currency.

```r
jobs <- c("data analyst",
          "data scientist",
          "data architect",
          "data engineer",
          "data analitics",
          "data reviewer",
          "business analyst",
          "data scpecialist",
          "database analysis",
          "data analysis",
          "business process analyst",
          "business systems analyst")
dat2 <- dat %>%
  filter(Currency=="USD") %>%
  filter(tolower(`Job title`) %in% jobs | tolower(`Job title - additional context`) %in% jobs | tolower
nrow(dat2)
```

```
## [1] 135
```

After searching for relevant jobs title or functional descriptions that contain phrases associated with data practitioners, there are 135 employees. Which represents 0.008 of the original dataset.

```r
summ <- data.frame(Median = median(dat2$`Annual salary (gross)`),
                   Mean = mean(dat2$`Annual salary (gross)`))
dat2 %>% ggplot(aes(x = `Annual salary (gross)`)) +
  geom_density(fill = "cornflowerblue") +
  theme_bw() +
  labs(title = "Distribution of Salaries for Data Practitioners", x = "Gross Annual Salary(USD)", y = "
  scale_x_continuous(labels = scales::label_comma()) +
  scale_y_continuous(labels = scales::label_comma()) +
  geom_vline(data = summ, aes(xintercept = Median, color = "Median"), linetype = "dashed", size = 1.25)
  geom_vline(data = summ, aes(xintercept = Mean, color = "Mean"), linetype = "dashed", size = 1.25, sho
  scale_colour_manual(name = "Statistic", values = c(Median = "darkred", Mean = "orange"))
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```


```
summ
```

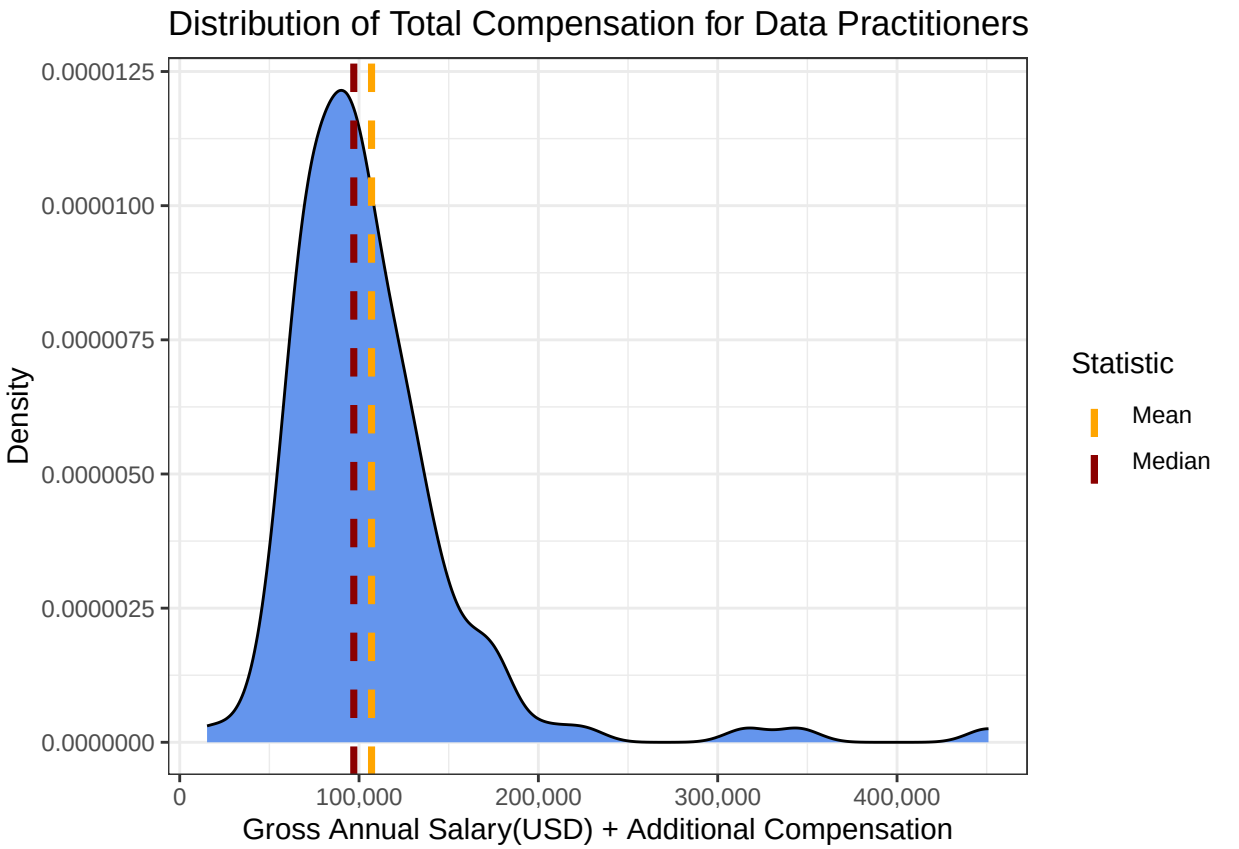
```
##   Median     Mean
## 1  93460 100732.1
```

The density plot shows a largely symmetric distribution with a mean $100,732(orange line), and median of $93,460. This according to this survey of 135 data practitioners the median salary is nearly $10k below the US Bureau of Labor Statistics 2022 estimate.

However, the dataset contains a column for additional compensation: **Does this increase the annual salary for data practitioners?**

```
dat2$total_salary = rowSums(dat2[,7:8], na.rm = TRUE)

summ <- data.frame(Median = median(dat2$total_salary),
                   Mean = mean(dat2$total_salary))
dat2 %>% ggplot(aes(x = total_salary)) +
  geom_density(fill = "cornflowerblue") +
  theme_bw() +
  labs(title = "Distribution of Total Compensation for Data Practitioners", x = "Gross Annual Salary(USD
  scale_x_continuous(labels = scales::label_comma()) +
  scale_y_continuous(labels = scales::label_comma()) +
  geom_vline(data = summ, aes(xintercept = Median, color = "Median"), linetype = "dashed", size = 1.25)
  geom_vline(data = summ, aes(xintercept = Mean, color = "Mean"), linetype = "dashed", size = 1.25, sho
  scale_colour_manual(name = "Statistic", values = c(Median = "darkred", Mean = "orange"))
```



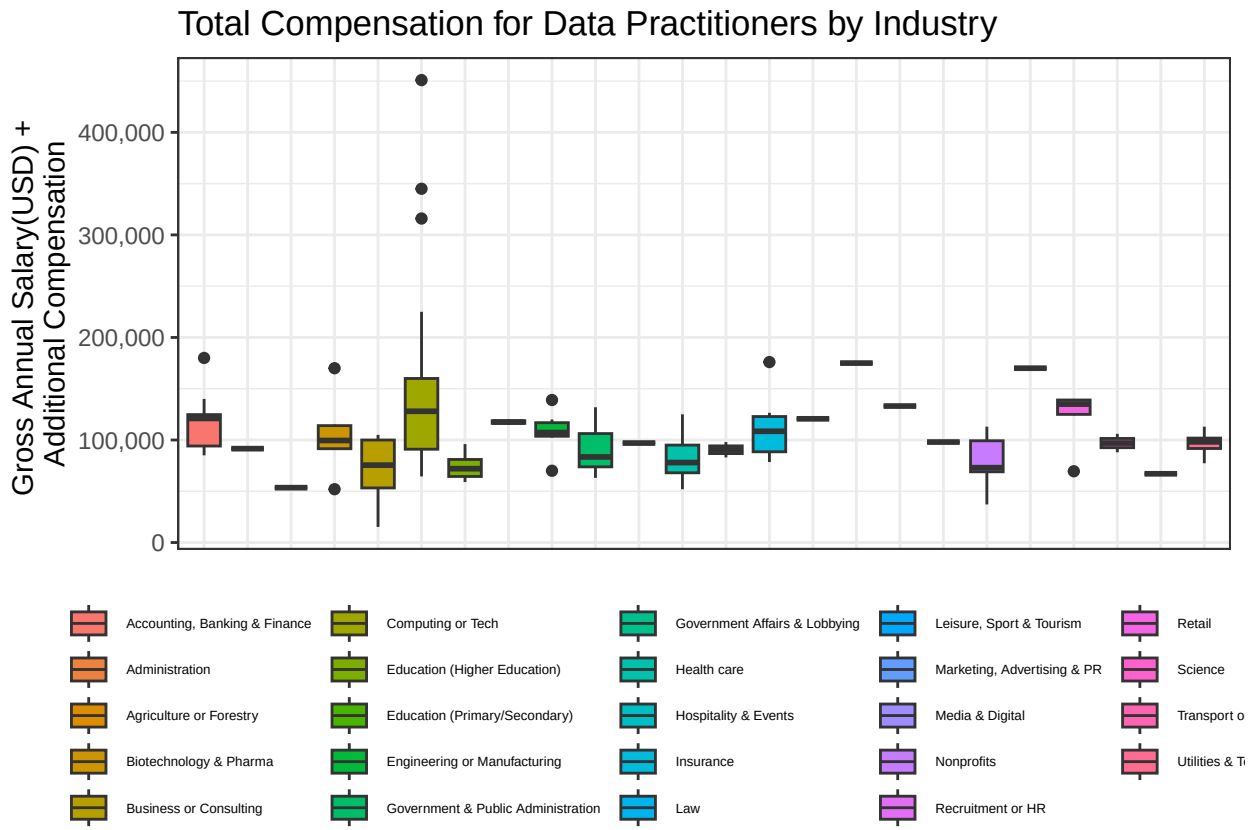

```
summ
```

```
##   Median     Mean
## 1  97060 106969.3
```

After including the additional compensation, the median total salary increased slightly to $97,060 and the mean increased to $106,969. The median is still below the estimate from the US Bureau of Labor Statistics.

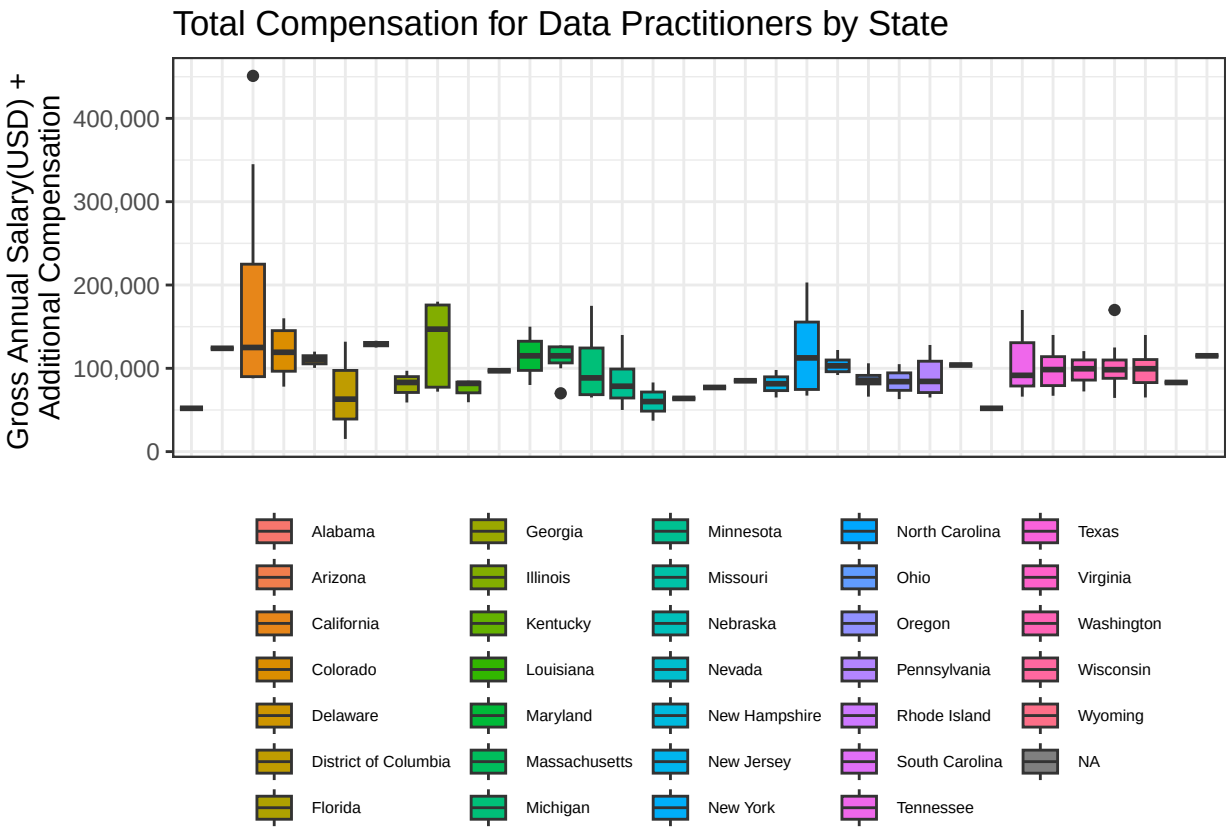
Which industries offer the highest compensation for data practitioners?

```
dat2 %>% ggplot(aes(y = total_salary, x = Industry, fill = Industry)) +
  geom_boxplot() +
  theme_bw() +
  labs(title = "Total Compensation for Data Practitioners by Industry", y = "Gross Annual Salary(USD) +
  scale_y_continuous(labels = scales::label_comma()) +
  theme(axis.ticks.x = element_blank(),
        axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text = element_text(size = 5),
        legend.title = element_blank())
```



Which States offer the highest compensation for data practitioners?

```
dat2 %>% ggplot(aes(y = total_salary, x = State, fill = State)) +
  geom_boxplot() +
  theme_bw() +
  labs(title = "Total Compensation for Data Practitioners by State", y = "Gross Annual Salary(USD) + \n
  scale_y_continuous(labels = scales::label_comma()) +
  theme(axis.ticks.x = element_blank(),
        axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text = element_text(size = 6),
        legend.title = element_blank())
```



CONCLUSION

The median and mean salaries for data practitioner and related fields according to the Google Sheets survey fall slightly shorter than that from the mean and median salary data we reviewed from the US Bureau of Labor Statistics. Further, when considering total salary which may includes other than salary compensation data, is only marginally more than the of US Bureau of Labor Statistics. Unsurprisingly states such as California and New York total compensation is some of the highest in the United States.