Analyzing Compensation for Data Scientists

Anmol Srivastava, Juan Solorio, Matthew Rhodes, and Andres De La Fuente

Abstract

The goal of this project is to evaluate the effects of certain factors (such as job title and education) on observed differences in salary for working data scientists. The analysis employs techniques such as Welch t-tests, ANOVA, and regression, performed on data provided by Kaggle's 2017 "ML and DS Survey" (https://www.kaggle.com/kaggle/kaggle-survey-2017). The factors explored in this project are physical location (specifically, whether the respondent is in a 'high density' or 'low density' area), job title, recommended programming language, and education. We found evidence (insert p-value) supporting the notion that location determines differences in data scientists' salaries. We did not find sufficient evidence to claim that programming language determines such differences, even within individual job titles. However, we did find evidence for a significant difference in salaries between different job titles. A positive relationship betwen education level and salary was similarly supported by the data.

Introduction

The focus of this project is compensation within the industry of data science. As prospective employees in the field, we have an interest in which factors might affect current data scientists' pay. Our analyses are built upon a dataset from Kaggle, which itself is the result of an industry-wide survey conducted on people working in the data science arena. This data is further explored in the 'Dataset Description' section. Then, after some exploratory analysis, we derived three questions upon which to center our research.

Question 1: Do data scientists' salaries differ between densely-populated and sparsely-populated areas?

Question 2: Do data scientists' salaries differ based on their job title? Do these salaries differ based on the programming languages recommended by respondents?

Question 3: Do data scientists' salaries differ based on the level of education they have attained?

For each of these questions, we aimed to test for both the presence of significant differences between groups, and for more specific relationships (via regression).

Dataset Description

As previously mentioned, the dataset from which we drew our conclusions is Kaggle's 'ML and DS Survey' for 2017. Kaggle's subsequent results are summarized as follows: "For the first time, Kaggle conducted an industry-wide survey to establish a comprehensive view of the state of data science and machine learning. The survey received over 16,000 responses and we learned a ton about who is working with data, what's happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field." The survey asked an extremely broad scope of questions, which resulted in a mixture of multiple choice responses (e.g. "Education Level") and freeform responses (e.g. "Best description of undergraduate major?").

The dataset is structured as follows:

schema.csv: A .csv file with the survey schema. This schema includes the full, exact questions that correspond to each column name in both the multipleChoiceResponses.csv and freeformResponses.csv sheets.

multipleChoiceResponses.csv: Respondents' answers to multiple choice and ranking questions. These are non-randomized and thus a single row corresponds to all of a single user's answers.

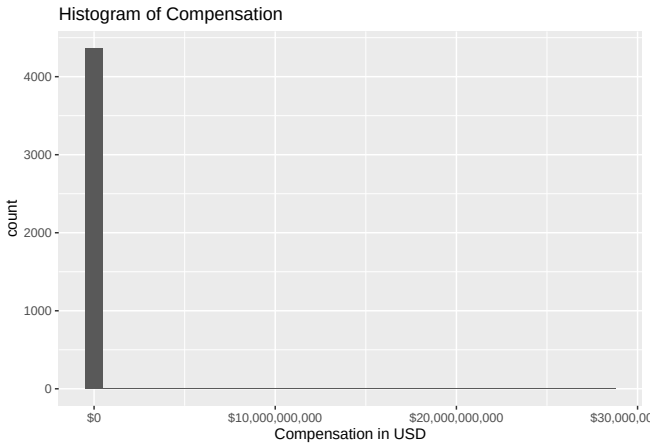
freeformResponses.csv: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading across a single row does not give a single user's answers.

conversionRates.csv: Currency conversion rates (to USD) as accessed from the R package "quantmod" on September 14, 2017.

RespondentTypeREADME.txt: This is a schema for decoding the contents of the schema.csv file.

We used the values in the dataset's conversion rates file to generate compensation values in USD for all respondents. For the purposes of our analyses, we limited our focus to multipleChoiceReponses.csv.

The following is a simple histogram of our generated compensation data (in USD) from the survey, before any kind of manipulation.

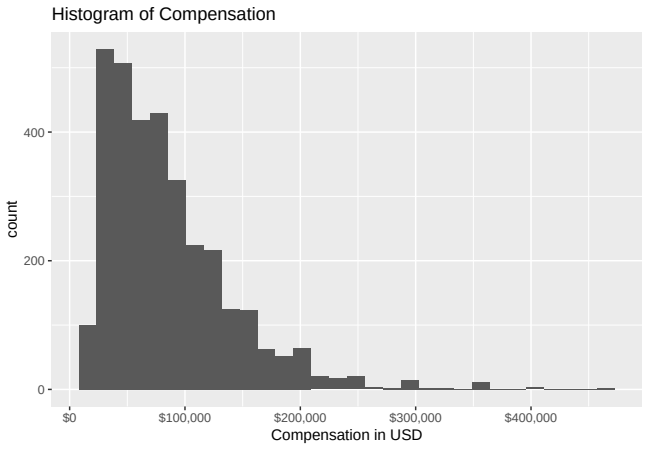


The mean for this uncleaned data is 6.605993×10^6 with a maximum value of 2.82974×10^{10}. Further exploration revealed that the dataset also contains a large quantity of zeros and values that seem too low to be someone's full-time pay. Based on our domain knowledge, we judged that these results were unrealistic, and therefore manipulated the data in a few important ways before applying our analyses.

First, we removed all entries with a zero for compensation, as this would either imply unemployment, entry error, or plain junk data, none of which were wanted for our analyses.

Second, we dichotomized compensation into two groups: 'salaries', and 'commissions' (the values which we decided are too low to be salaries). We decided to split the values at 20K; anything below this is considered commission.

Third, we decided that some values on the high end were either entry error or extreme outliers (there were values ranging from millions to billions of dollars), and should be capped. The capping value was set to $500,000 based on our real world experience. As evidenced below, the resulting dataset was much more reasonable to base further work on.



A secondary dataset was employed to help answer Question 1 (https://population.un.org/wpp/Download/Standard/Population/). The dataset is described as follows:

Total annual population, by sex, from 1950 to 2100.

```
PopMale: Total male population (thousands)
PopFemale: Total female population (thousands)
PopTotal: Total population, both sexes (thousands)
PopDensity: Population per square kilometre (thousands)
```

We were only interested in the lastest population, so we used the entries for 2019. After we had these values, they were converted in population per square mile for interpretability. The goal of this data was to allow for a determination of high-density and low-denisty countries. All of the countries that were included in the multipleChoiceReponses dataset were also included in the population dataset, so we appended the appropriate densities to their respective rows in the multipleChoiceReponses dataset as a final step before starting our analyses.

Statistical Methods and Assumptions

ANOVA:

The first method we arrived at was Analysis of Variance (ANOVA). From lecture, we know that ANOVA is designed to provide a single test of a null hypothesis of equal group means with a desired significance level. It is a generalization of the equal-variance t-test to the case where the number of means to be compared is greater than 2. We are testing for difference in means of various groups, hence using ANOVA would be convenient and efficient.

Regression

The second test we thought to use was linear regression. From lecture, we know that linear regression is equivalent to ANOVA when the variances are equal, and the null hypotheses for ANOVA and regression both imply that the mean response does not depend on the predictor. We also ackowledge that ANOVA and regression will not always agree, which is why we decided to run both.

Welch T-Test

After some exploratory analysis, we found that sample variances showed notable differences, meaning we might not necessarily satify the pertinent ANOVA assumptions for each group. This led us to conduct a Welch T-Test as well, addressing the differences in variances, and thus the issues that might arise from the use of ANOVA.

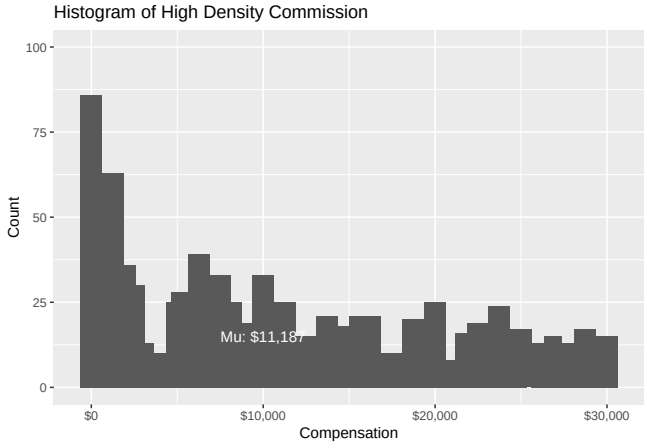
Assumptions: For the each of the tests, we are working under the assumption that the samples included in the Kaggle Survey dataset are independent, meaning that no answers provided by one surveyee affected another surveyee's entry. Moreover, we are working under the

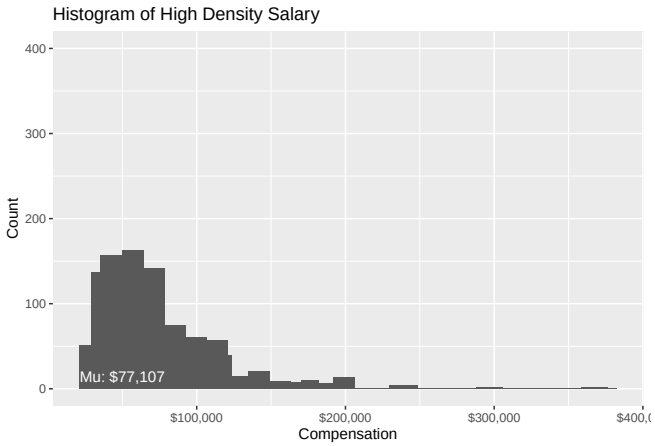
assumption that we have sufficiently large data sizes for each group, to account for the normality requirement in both the ANOVA and Welch tests. For the ANOVA and regression models, we are also assuming equality in the variance of the groups. Finally, we presume normality and liniarity for the models, to construct either linear or generalized linear models.

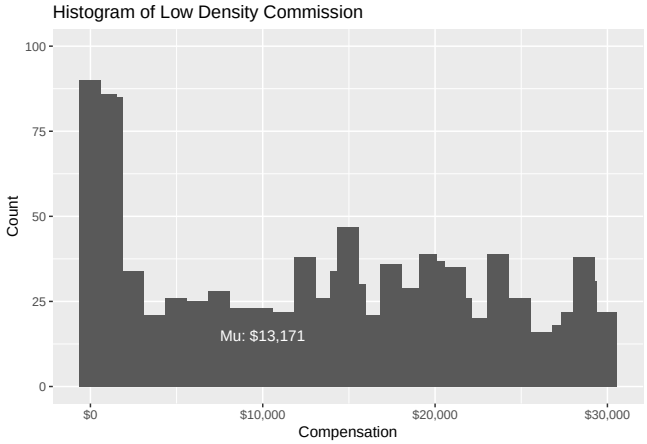
Pay & Location

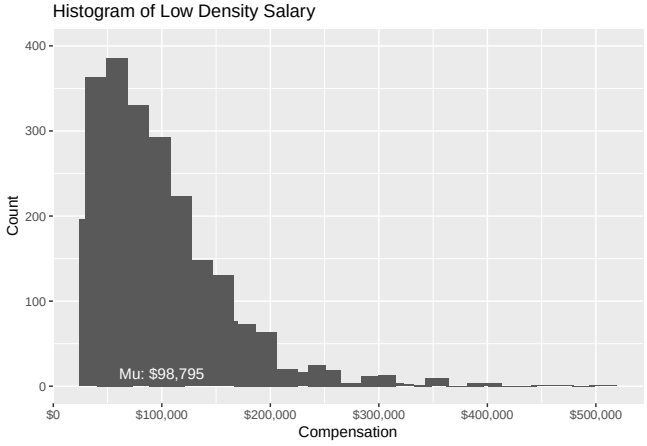
For question 1, prior to conducting our experiment, we anticipated that the amount of information provided by the dense areas will be significantly larger than the data provided by sparse areas. We are assuming this because we think that areas with a high population will have more of a need for data scientists than sparse areas (like rural countries). One important assumption we are making (which affected our choice of method) is that the data is normally distributed. This means that areas that are on the lower end of sparse and dense will appear as often as areas that are on the higher end of dense and sparse, while the majority of areas fall closer to the mean of each population density. We initially felt that this would be a problem. However, even if the data wasn't normally distributed, since we have around 10,000 rows and roughly 4 features, it is believed that the sample size is sufficiently large. Lastly, as previously stated, we make the equal-variance assumption.

Below are histograms for commission and salary, broken down into high density and low density areas.

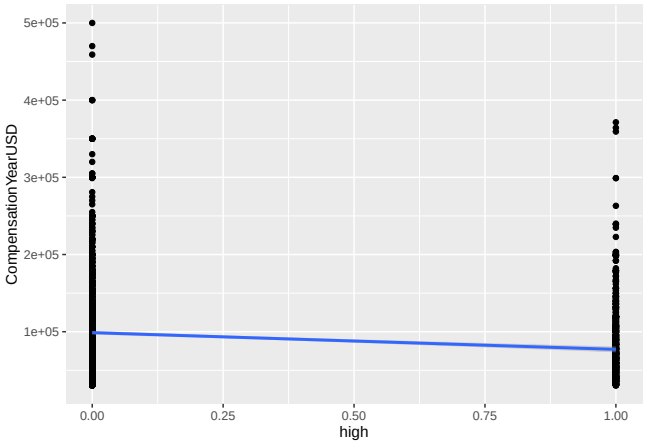


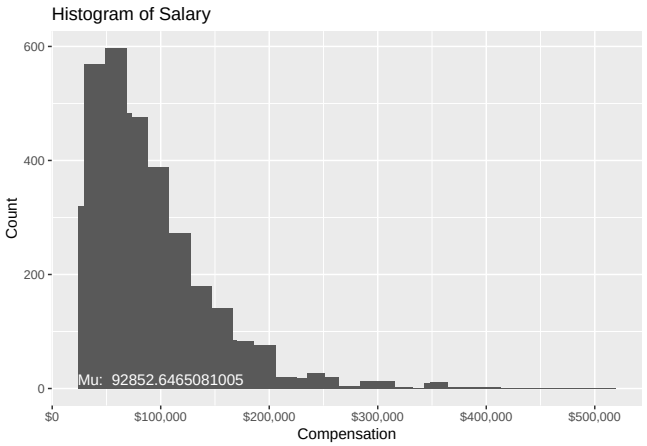




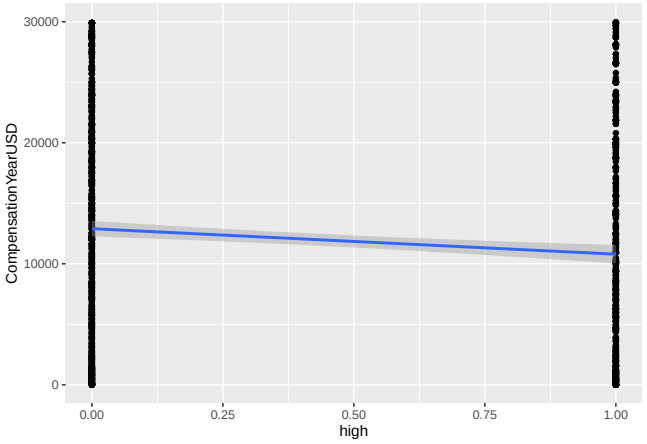


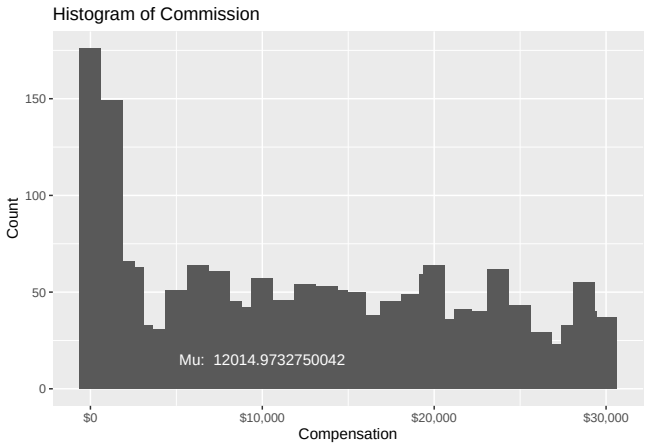
From linear regression tests, there is a slightly negative linear relationship between density and compensation (for salary and commission). The x-axis denotes 1 for high and 0 for low density, so we can see that (as density goes from low to high), average salary goes down by $21,652.





This is the linear regression plot and the distribution of the total commission (including high and low density). Similarly, as density goes from low to high, average commission goes down by $2,106.





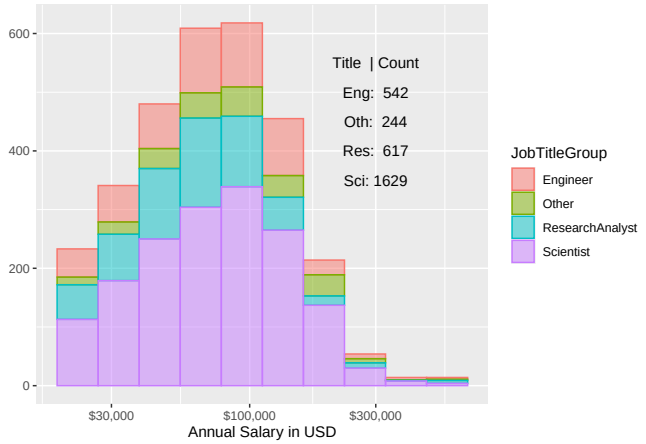
Pay & Job Title / Programming Language

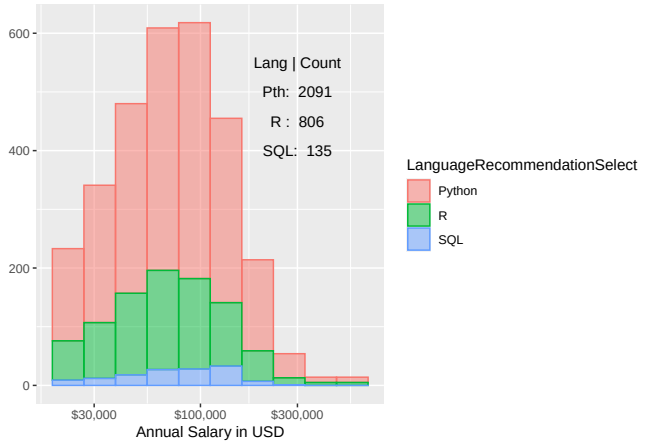
For question 2, we wanted to explore the effects that a specific job title or preference in programming language might have on salary. For our purposes, our questions narrowed down to:

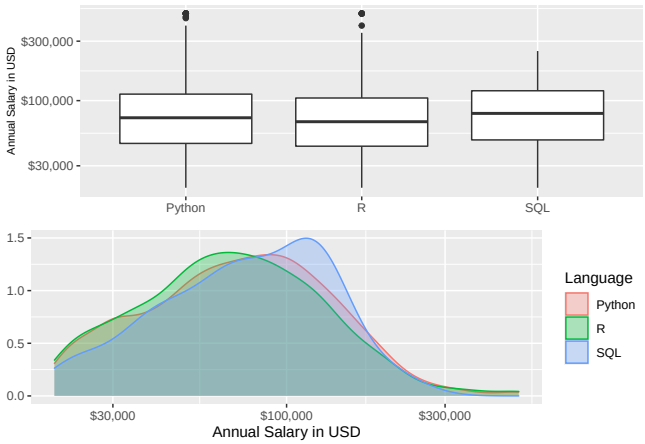
Are data scientists' salaries affected by job-title group (i.e. Data Analyst/Data Scientist, Engineer, Researcher), or by the type of programming language they employ most?

Within the original survey, the participants were provided with 17 different titles as options for that which best suited their current role. To answer the first question we needed to create groups for the 17 different entries to the 'Current Job Title' descriptions. For this we decided upon grouping titles into 'Scientist' (people probably utilizing higher level math/statistics), 'ResearchAnalyst' (those focused with general data analysis and manipulation), 'Engineer' (titles focused on data engineering or software), and 'Other' (description unknown). We worked with the already-capped salary data for this question.

Additionally, while there is no entry for "programming language participant most frequently employs", surveyees were given the free-hand option of 'Language Recommendation' they would give to people wanting to enter their field, which we used as a proxy. There were 14 languages (ex: Java, Python, R, Scale, etc), of which we were only interested in Python, R, and SQL (these were the top 3 recommendations). This narrowed our original dataset of over 16,000 rows, to one of around 3,000 for this part of the project, distributed as shown here:







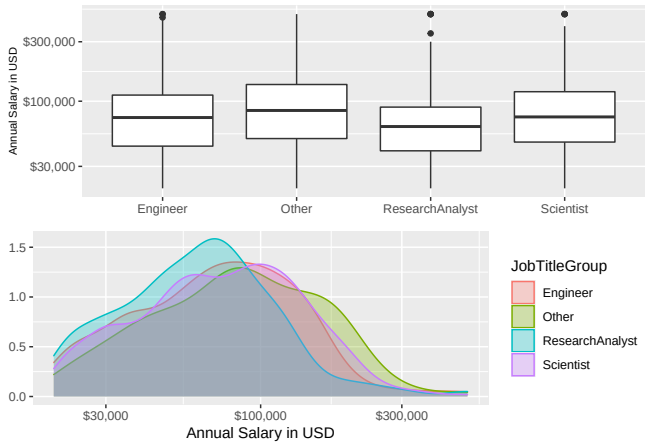
The boxplots for both the job title groups and language preference groups seem to have minimal difference in their variance. The density plots for each group also suggests a general normality in the distribution of the groups. Given these observations, we proceeded with the needed assumptions of equal variance and normality for ANOVA and regression.

For our tests, we ran under the hypotheses:

$$H_0 : There\ is\ no\ difference\ in\ mean\ salary\ (\mu_1 = \mu_2 = ... = \mu_n)$$

$$H_1 : There\ is\ difference\ in\ mean\ salary$$

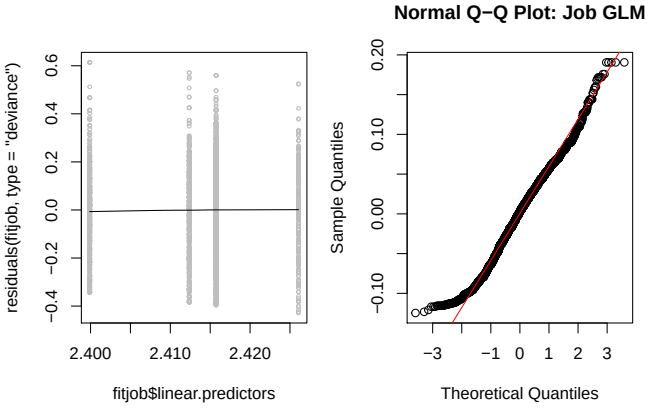
Working with this subset for job titles and programming languages, we had satistifed the need for a sufficiently large dataset, but we still needed to make sure our assumptions of equal variance and normality were held to then apply our tests. The data was passed into some box-plots and density plots to check for the variance and normality of the distribution as shown in the figure here:

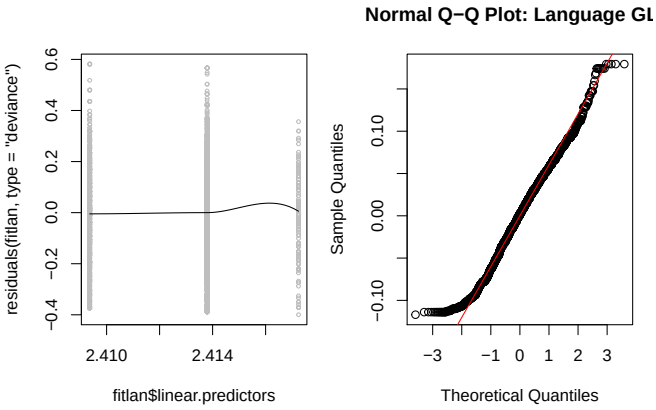


Going with the assumption of a normal distribution for the Capped Salaries, the Type 1 error probability of the ANOVA test was calculated to be p-val=0.042 through MC-simulations. This falls within the critival value of 0.04-0.06 for our alpha.

From the ANOVA tests, when testing at the $\alpha = 0.05$ significance level, we found the job title groups to have a p-value = 9.8859e-09, yielding strong evidence to reject the null hypothesis and implying that the mean salary of the job title groups are not equal. On the other hand, for the language groups we have the p-value = 3.4716e-01, meaning we do not have strong evidence to reject the null hypothesis of equal mean salary among language preferences.

We further test the null hypothesis through a generalized linear model (glm), log-linear regression, as this might provide a different conclusion than the ANOVA test. The log-linear model was chosen because it would return a ratio relative difference in the mean Salary for the groups compared to the Scientist groups. This would be a better way to express the difference in Salary instead of the direct value comparison of the mean Salary difference given by a linear regression model.





Pay & Education

Discussion

We first address our original assumptions about Question #1: we did not ultimately receive more data from the high density areas. In fact, we received more data from the low density areas. The data points from the dataset were indeed independent, since this survey was given to individual participants, but we did not have a normal distribution of data for commission or salary, nor equal variances for either of them. Subsequently, the initial metric statistics lead us to conclude that we should not use ANOVA or linear regression, as the variances are not the same. We also appreciate that unequal variances are a problem no matter how large the sample size may be, thus the Welch method is more appropriate. However, we also realize that there is a malleable window of difference (a certain thershold) that counts as equal variance, so we still performed our other tests.

Recall that commission is defined as anyone making less than $20k, all other respondents are classified as salaried. The p-value reported from the anova tests are <2e-16 for salary and .00011 for commission. The p-value for salary and commision are 2.2e-16 and 0.0001096, respectively, so there is remarkable evidence to reject the null hypothesis. Consequently, we conclude that density does have an effect on compensation (both salary and commision). There is defintely some existing difference, and if we inlcude the results from our linear regression model, we can further point towards a possible a negative association between density and compensation.

Question 2:

Question 3:

limits of analyses: Question 1: Unequal variances.

Question 2:

Question 3:

limits of data too: (I think this applies to all questions) Question 1:

Question 2:

Question 3: ~~~ talking abt stuff in context of references ~~~

We believe our work has strong connections to workforce-oriented research in the real world. Individuals have come across conclusions that are similar to ours, albeit with varying degrees of agreement. For instance, a finance website hosts an investigative piece on the importance of education level in data sience. It deviates mildly from our findings (identifying the relationship between degree and pay as mere positive correlation), but stresses that education level explains workers' skillsets and experience – which may then translate to career and salary growth (Hayes). Similarly, a U.S. Bureau of Labor Statistics study cofirms our suspicions that location matters a great deal – going so far as to claim that factors like cost of living or pace of life may supercede other attributes as determinants of compensation for identical work (Torpey).

In any case, our project has the potential to benefit not only prospective data scientists, but also to inform the industry at large as to the state of affairs in the data science arena. Our tests (and additional tests using the unexplored data in the complete survey) could continue to be used in this regard, to unearth even more surprising and beneficial patterns in this ever-growing field.

References

Hayes, Bob. *"When Does Education Level Matter in Data Science?"* Business Broadway, 2020. (https://businessoverbroadway.com/2016/03/14/when-does-education-level-matter-in-data-science/)

Kleiman, Iair. *"Data Scientists' Salaries Around the World V2.0."* Kaggle, 2017. (https://www.kaggle.com/kaggle/kaggle-survey-2017)

Torpey, Elka. *"Same Occupation, Different Pay: How Wages Vary."* U.S. Bureau of Labor Statistics, 2015. (https://www.bls.gov/careeroutlook/2015/article/wage-differences.htm?view_full)