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 relevance of certain factors (such as job title and education) towards observing differences in salary for working Data Scientists. The analysis employs techniques such as Welch t-tests, ANOVA and Regression, and is based on the dataset 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, suggested programming language, and education. We found evidence (insert p-value) that physical location indicates a difference in the salary of a Data Scientist. We did not find strong evidence that programming language indicates a difference, even within individual job titles. However, we did find strong evidence for a difference in salaries between job titles. We also found evidence for a positive relation between education and salary.

Introduction

The focus of this project is compensation within the industry of Data Science. As future employees in the field, we have an interest in which factors might relate to current Data Scientists' pay. To base our analysis of this topic, we decided on a dataset from Kaggle, which is the result of a wide survey conducted on people working in the field (https://www.kaggle.com/kaggle/kaggle-survey-2017). See the next section for our description of the data.

After some exploratory analysis, we came up with three questions to focus our efforts on.

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 salaries differ based on the programming languages an individual recommends?

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 differences between groups, and for more specific relationships by using regression.

Data Set Description

As referenced above, the dataset we utilized to draw our conclusions is Kaggle's 'ML and DS Survey' for the year 2017. Kaggle's description of the dataset reads 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 dataset with a mixture of multiple choice responses (e.g. "Education Level") and free form responses (e.g. "Best description of undergraduate major?").

The dataset contains the following:

schema.csv: a CSV file with the survey schema. This schema includes the questions that correspond to each column name in both the multipleChoiceResponses.csv and freeformResponses.csv.

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.

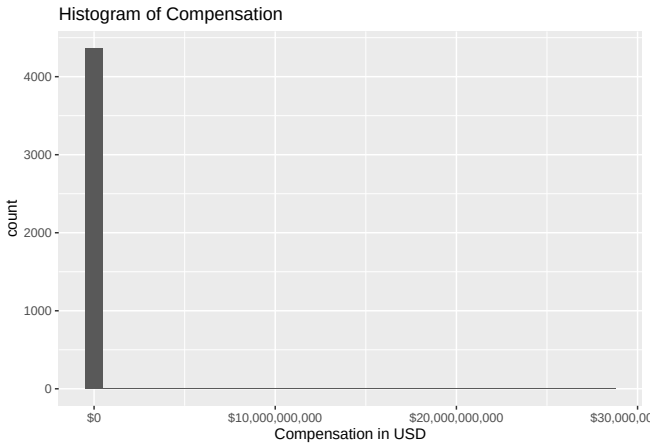
freeformResponses.csv: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading across a single row does not give a single user's answers.

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.

Below is a simple histogram of our generated compensation data in USD from the survey before any kind of manipulation.

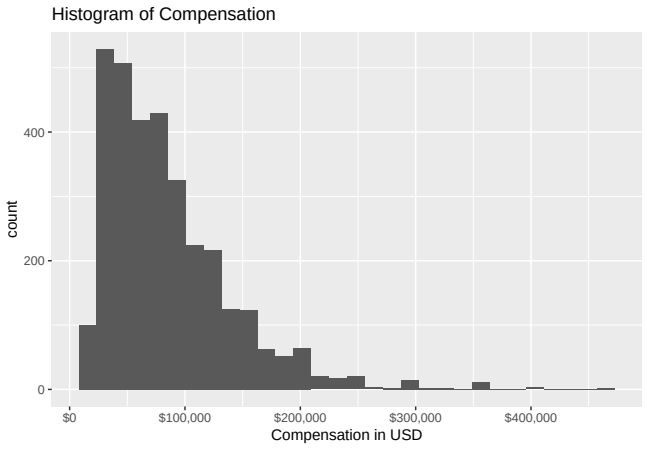


The mean for this uncleaned data is 6.605993×10^6 with a maximum value of 2.82974×10^{10}. Further exploration of this revealed that the dataset also contains a large quantity of zeros and values that seem too low to be someone's full time pay. Based on our domain knowledge of the subject at hand, we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

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 data set 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 population by sex, annually 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)
```

First, 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 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 we started our analyses.

Statistical Methods and Assumptions

ANOVA:

The first method we thought to use was Anova. From lecture we know that Analysis of Variance 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, hense using ANOVA would be very convenient and efficient.

Regression

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

Welch t-test

After some exploratoy analysis of the data we found that the variances showed some difference, hense we might not be able to satify assumptions for ANOVA for each group. This lead us to conduct a Welch T-Test as well, addressing the difference in variance, addressing the issues that might arise from ANOVA.

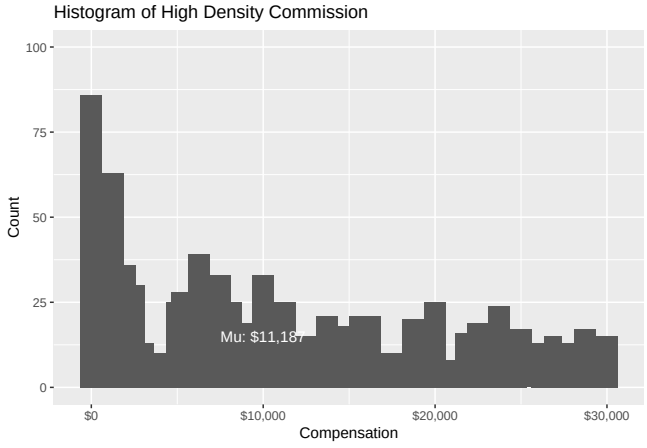
Assumptions: For the each of the test we are working under the assumption that the samples included in the Kaggle Survey dataset are independent, meaning that no

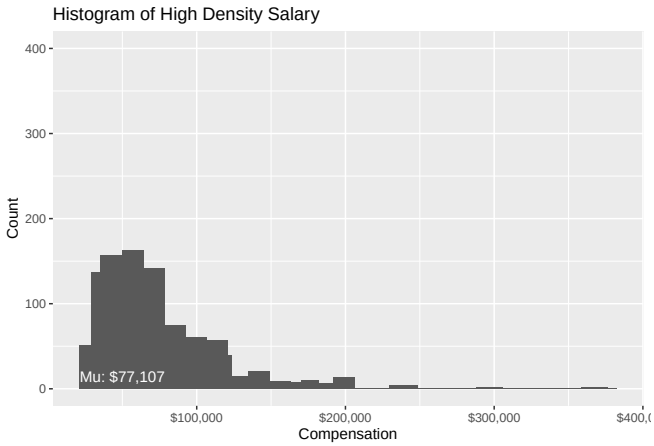
answers provided by someone that took the survey affected a different entry. Moreover, we are working under the assumption that we have large enought data for each group to account for normality in both the ANOVA and Welch t-test. For the ANOVA and Regression models, we are also assuming equality in the variance of the groups. Finally, we persume there to be normality and liniarity for the models to perform either Linear or a Generalize linear model.

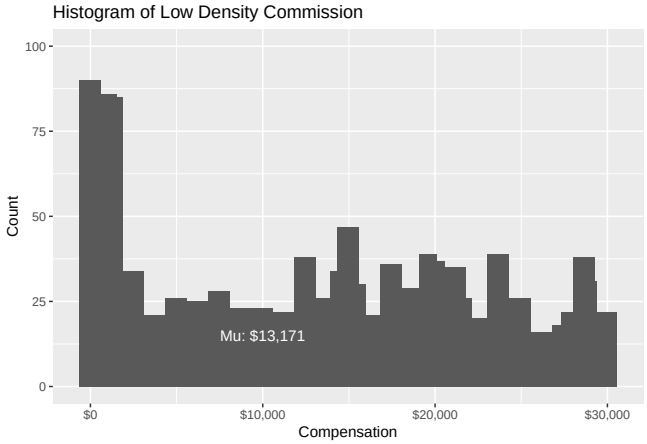
Pay and Location

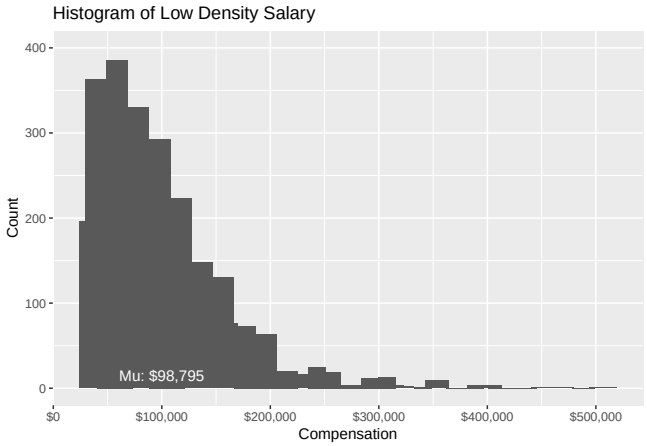
For question 1, Prior to conducting our experiment some of the things that we were assuming going into this question is that the amount of information provided by the dense areas will be significantly larger than the data provided by sparse areas. We are assuming this because we think that areas that have a high population will have more of a need for data scientists than sparse areas like rural countries. Another assumption that we are making is that each one of the samples that are included in the dataset are independent, meaning that no answers provided by someone that took the survey affected someone else's response. One important assumption we are making which affected the method we want to use is that we have a normal distribution of data. 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 thought that this would be a problem.Even if our data wasn't normally distributed, since we have around 10,000 rows and roughly 4 features we think the sample size will be large enough. The last thing we are assuming for this is that there is Equal variance.

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

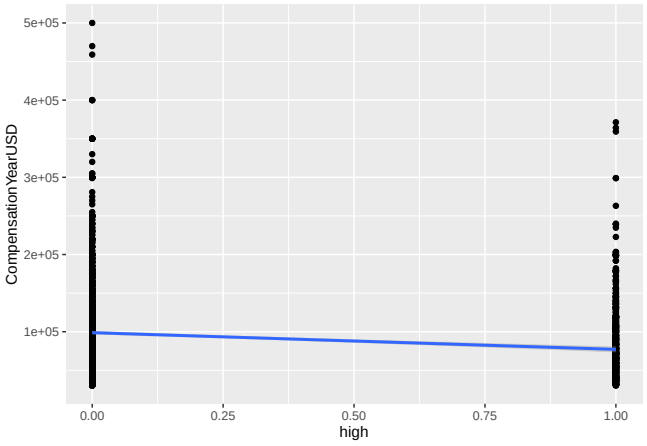


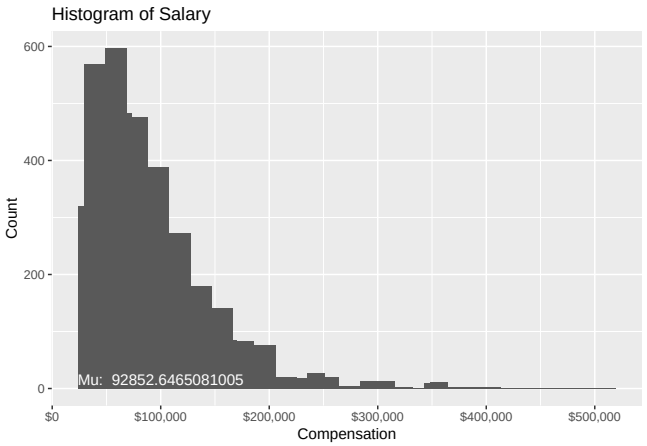




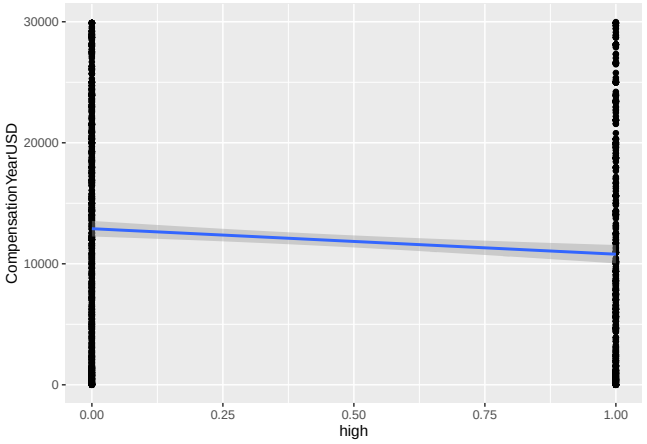


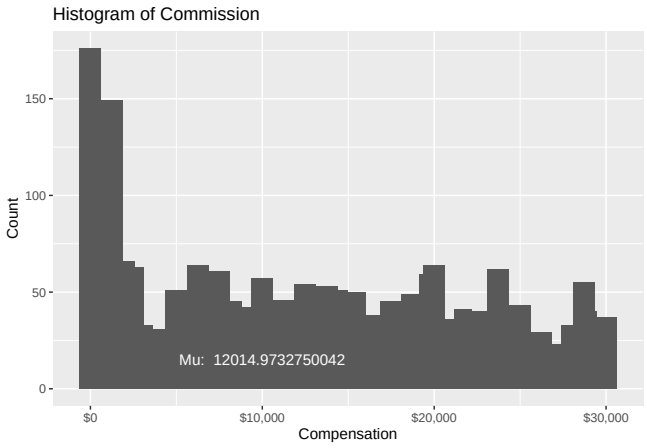
From Linear Regression, there is a slight negative linear relationship between density and compensation (for salary and commission). X axis is 1 for high and 0 for low density, so we can see that as density goes from low to high on average Salary goes down by $21,652.





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 on average commission goes down by $2,106.





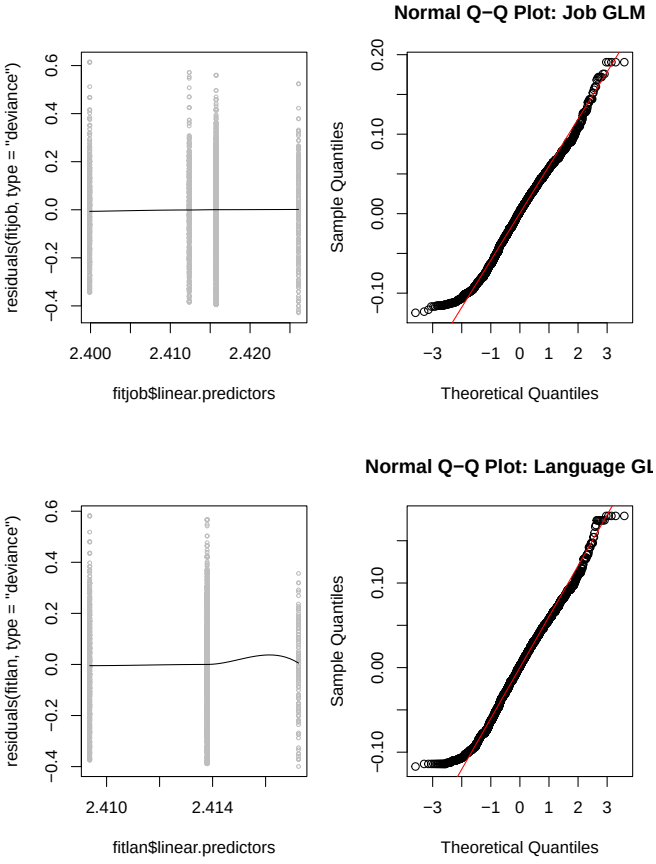
Pay and Job Title / Programming Language

In question 2, we wanted to explore the effects that a specific Job Title or Preference in Programming Language might have on salary. For our purpose our questions narrowed down to:

Are Data Scientist salaries affected by the Job Title Group (i.e. Data Analyst/Data Scientist, Engineer, Researcher) or by the type of Programming Language they employ most?

Within the original data, the participants were provided with 17 different Job Titles as options for title that best suit their current position. 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 in the data engineering or software), and 'Other' (description unknown). We worked with the already capped salary data to answer the question. This lead our original dataset of over 16,000 to one of over 3,000 for this question, distributed as described in the figures below:

For the second part of the question, while there is no entry for "Programming Language Participant most employs", they 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 as these were the top 3 recommendations.



Pay and Education

Discussion

summarize results and conclusions: Question 1: To adress our original assumptions about this question we did not receive more data from the high density areas we actually received more data from the low density areas. The data points from the dataset were independent since this survey was given to individual participants, but we did not have a normal distribution of data for commission or for salary or equal variance for either of them. Conclusion from the initial metric statistics, is that we should not use Anova or linear regression because the variances are not the same, and we know that unequal variances are a problem no matter how large the sample size thus we should use Welch. However, we also know that there is a certain thershold that counts as equal variance so we ran the other tests anyway. Commission was define as anyone making less than $30k and otherwise they are classified as salaried. The p-value reported from the anova tests are <2e-16 for salary and .00011 for commission.

When looking at the p-value for salary and commision they are 2.2e-16 and 0.0001096 respectively so there is evidence to reject the null hypothesis meaning that density has an effect on compensation for salary and commision. We know that there is defintely a difference and if we inlcude the results from our linear regression model, we can conclude that there is 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:

References

brief

Appendices

more technical aspects of analyses, any other tidbits