Analyzing Compensation for Data Scientists

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

Abstract

The goal of this project is to determine whether there are differences in salary for working data scientists based on certain factors (such as location, job title and education). To that end, we employ techniques such as Welch t-tests, ANOVA, and regression 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 that data scientists' salaries differ based on location, job titles, and educational level. We further found evidence for a positive linear relationship between educational level and salary and a negative linear relationship between population density and salary. However, we did not find sufficient evidence for differences in salary based on programming language, even within individual job titles. We finish with a comparison of our findings with the work of individuals doing similar research.

Introduction

As prospective employees in the field, we have an interest in which factors might affect current data scientists' pay. For this reason we conducted our analyses on a dataset from Kaggle which is the result of a recent industry-wide survey conducted on data science professionals. This data is described in more detail in the following section. After some exploration of the data, we formulated three questions to guide our analysis.

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 their recommendation of programming language?

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 describes the dataset 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?").

Contents of the dataset:

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.

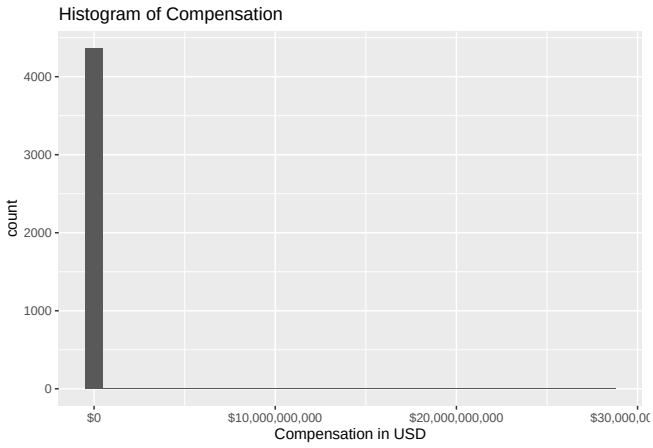
freeformResponses.csv: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading across a single row does not give a single user's answers.

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 calculated compensation data from the survey, before any kind of manipulation.

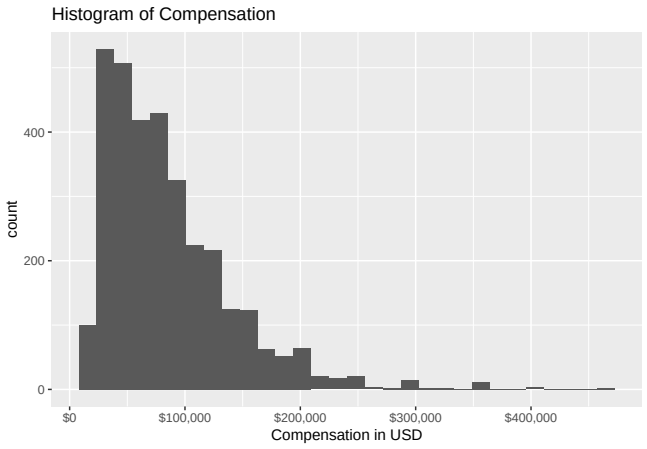


The mean for this uncleaned data is 6.605993×10^6 with a maximum value of 2.82974×10^{10}. Further exploration revealed that the dataset also contains a large quantity of zeros and values that seem too low to be someone's full-time pay. Based on our domain knowledge, we judged that these results were 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 eiher 20K or 30k; anything below the chosen threshold 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
PopFemale: Total female population
PopTotal: Total population, both sexes
PopDensity: Population per square km

(each of these is in thousands)
```

We were only interested in the latest 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 after all of the pre-processing was completed the amount of relevant data varied from question to question.

Statistical Methods and Assumptions

ANOVA:

The first method we decided to use was Analysis of Variance (ANOVA). ANOVA is designed to provide a single test of the null hypothesis that group means are equal 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. Since each of our questions deals with there being a difference between various groups, ANOVA is an obvious and efficient choice of method. However, there is a potential issue with our use of ANOVA which is discussed further in our concluding 'Discussion' section.

Regression

Linear regression was also used, though not for the same purpose in every question. In question 1 for example, it is a supplement to ANOVA and Welch T Tests in that it is detecting a difference between two groups. In question 3, however, it is used to detect a positive linear relationship. In the context of being used to detect a difference, linear regression is equivalent to ANOVA when the variances are equal. 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 the variances of the data groups formed for each question are not

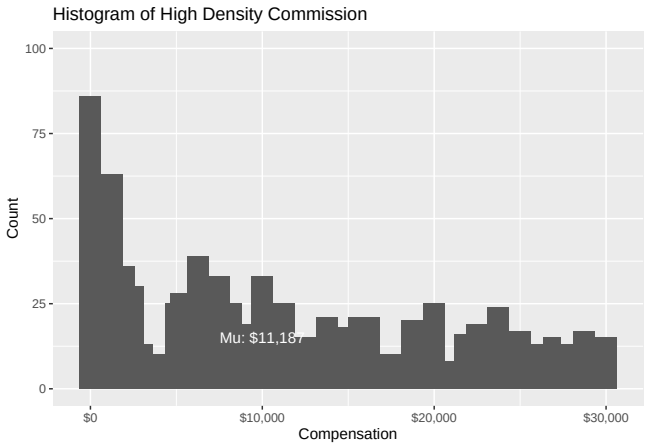
exactly equal. Since this could be a problem for ANOVA, we decided to conduct Welch T-Tests as well, addressing the differences in variances.

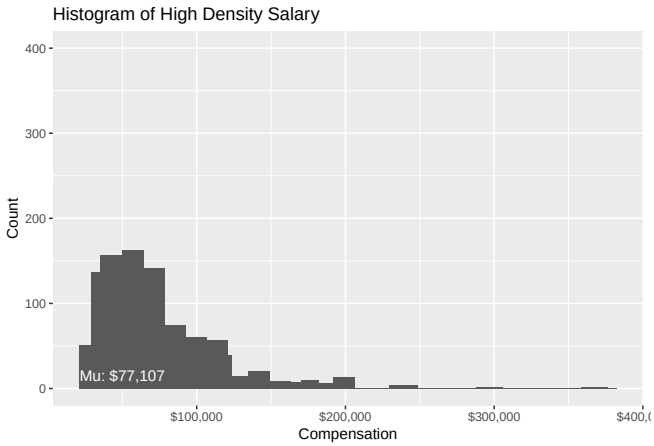
Assumptions: For the each of the tests, we are working under the assumption that the samples included in the Kaggle Survey dataset are independent, meaning that the answers of one respondent do not affect the expectation of answers for another respondent. Moreover, we are working under the assumption 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 assume the linearity of the relationship for the questions in which we applied linear regression to detect a relationship.

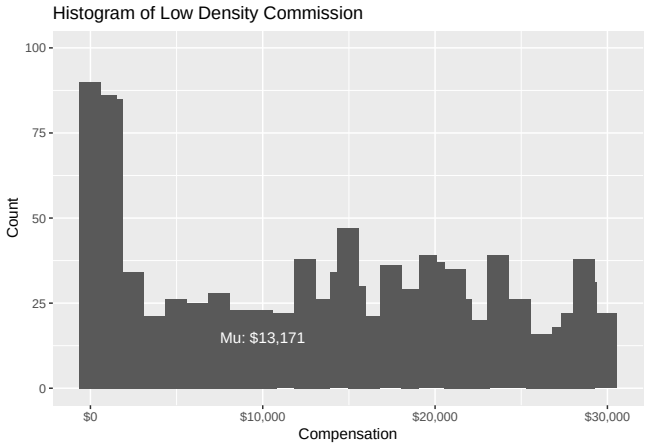
Pay & Location

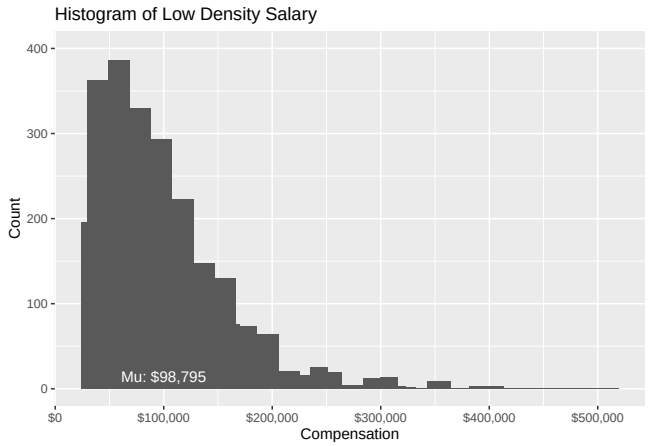
Prior to conducting our analyses into Question 1, we suspected that the amount of information from the dense areas would be significantly larger than the data from sparse areas. We thought this because of the assumption that areas with a high population will have more of a need for data scientists than sparse areas (like rural countries). One important assumption we are making (which affected our choice of method) is that the data is normally distributed. For this question, this means that areas that are on the lower end of sparse and dense will appear as often as areas that are on the upper end of sparse and dense, while the majority of areas fall closer to the mean of each population density. Even if the data wasn't normally distributed, since we have around 10,000 rows it is believed that the sample size is sufficiently large. The last thing we are assuming for this is that there is equal variance.

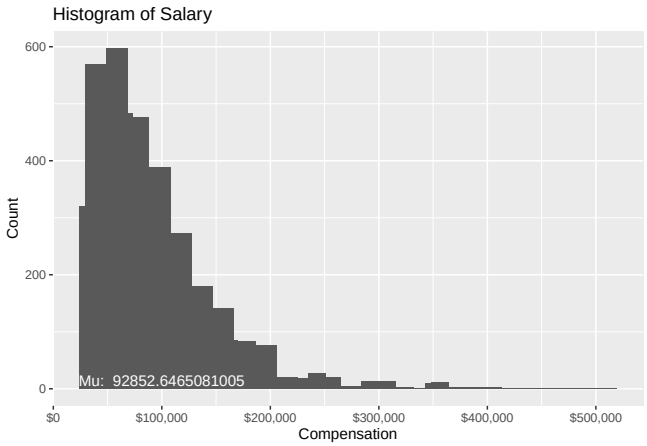
The data had a very wide range of values and we came to the conclusion that people are not making less than 20k-30k as a salaried employee. We decided to split the data into workers who make less than 30k as individuals who make commission and those that make more, as salaried here. High density areas are countries that have more than 500 people per square mile, while low density areas have less than 500. Below are histograms for commission and salary, broken down into high density and low density areas. As we can see from the histograms the data is not normally distributed so our initial assumption was incorrect. This distribution looks more exponential than normal based on our salary split.







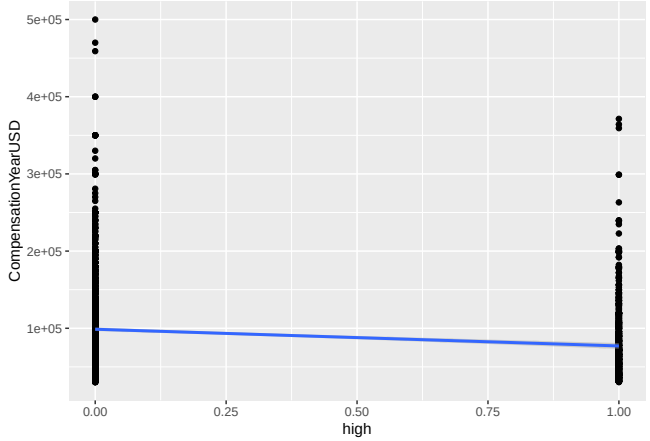




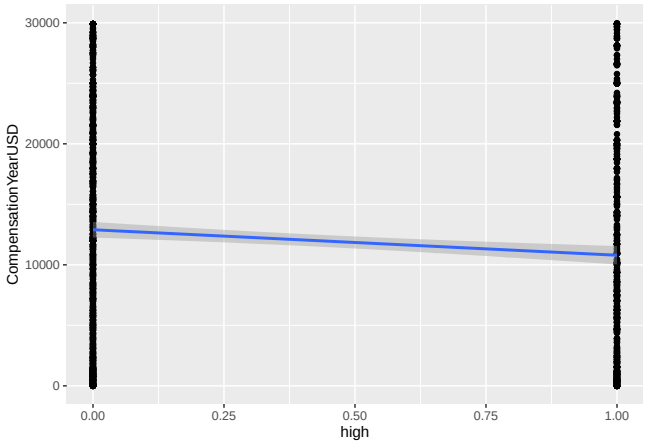
We then did power calculations to verify that we had an appropriate amount of power to conduct a valid experiment. The sample sizes we had after the preprocessing was 574 for high density commission and 805 for low density commission which allowed us to achieve a power of 97%. While sample sizes of 787 and 2098 for high and low salary allowed us to achieve a power of 99%.

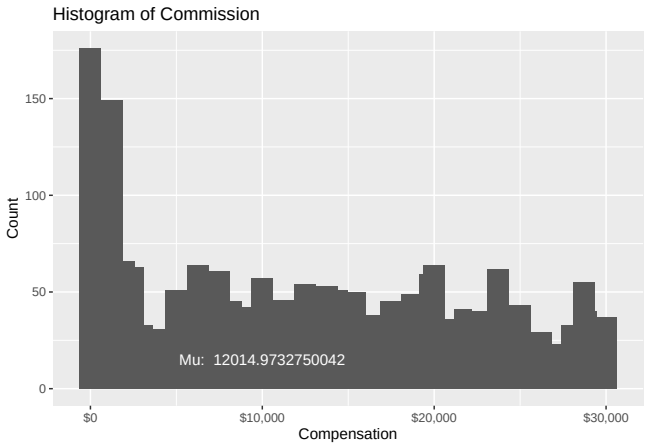
From linear regression tests, we observed that there is a slightly negative linear relationship between density and compensation (for salary and commission). In both levels of compensation (commission and salary), as your density goes from low to high your compensation decreases.

The 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. We also observed the distribution of the combined low and high density salaries.



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. We also observed the distribution of the combined low and high density commission.





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 understand 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.

Is there a differnce in pay between workers that live in high density and low density areas?

From our experiments we would say that there is defintely a difference in pay between these types of workers and if we inlcude the results from our linear regression model, we can further conclude a negative linear association between density and compensation. We concluded that this negative association exists most likely because low density might lead to more space between individuals, a higher standard of living, or bigger housing, which all might contribute to higher paying jobs. Our rational was that coutries that have a high density would have more people living in tightly packed areas that couldn't afford more expensive (bigger housing).

Recall that commission is defined as anyone making less than $30k, all other respondents are classified as salaried. The p-value reported from the anova tests are <2e-16 for high salary compared to low salary and .00011 for low commission and high commission. Since, the p-values for salary and commision are 2.2e-16 and 0.00011, respectively, there is remarkable evidence to reject the null hypothesis, but since we did not consider the variances to be equal we ran the Welch test as well. The p-value reported from the welch tests are <2e-16 for high salary compared to low salary and .00002765 between low commission and high commission. Consequently, we conclude that density does have an effect on compensation (for both salary and commision) and that ultimately there is a difference in pay between workers that live in high density areas compared to low density areas.

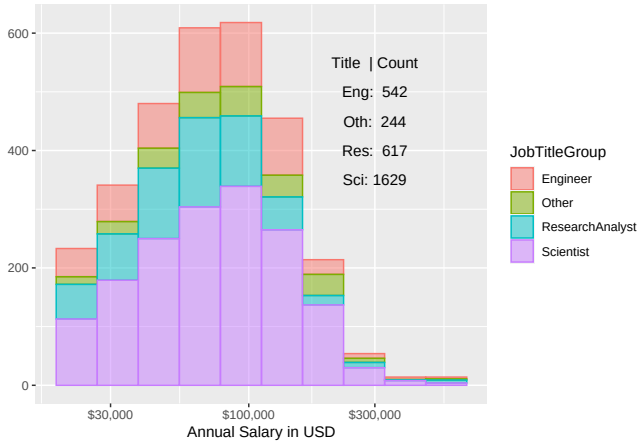
Pay & Job Title / Programming Language

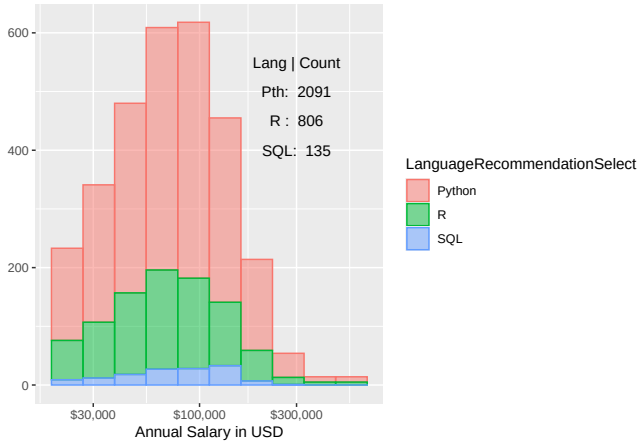
For question 2, we wanted to explore the effects that a specific job title or preference in programming language might have on salary. For our purposes, our questions narrowed down to:

Are there differences in data scientists' salaries when grouped by job-title (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.

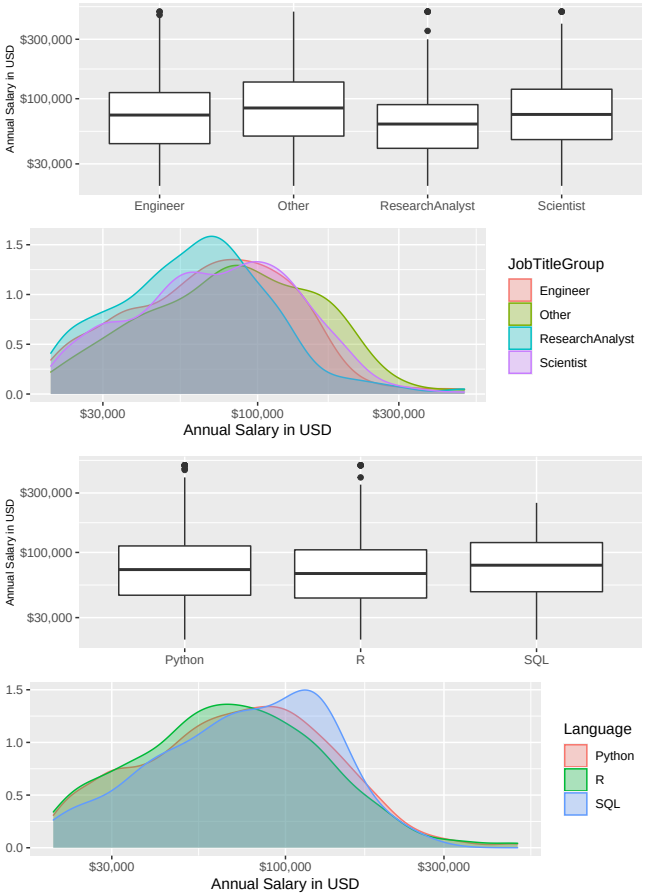
Additionally, while there is no entry for "programming language participant most frequently employs", surveyees were given the free-hand option of 'Language Recommendation' they would give to people wanting to enter their field, which we used as a proxy. There were 14 languages (ex: Java, Python, R, 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:





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:



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:

$$H_0 : There\ is\ no\ difference\ in\ mean\ salary\ (\mu_1 = \mu_2 = ... = \mu_6)$$

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

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.

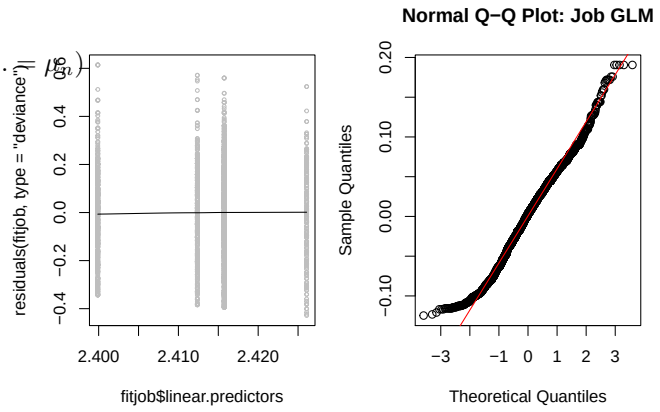
From the ANOVA tests, when testing at the $\alpha = 0.05$ significance level, we found the job title groups to have a p-value = 9.8859e-09, yielding strong evidence to reject the null hypothesis and implying that the mean salary of the job title groups are not equal. On the other hand, for the language groups we have the p-value = 0.347, meaning we do not have strong evidence to reject the null hypothesis of equal mean salary among language preferences.

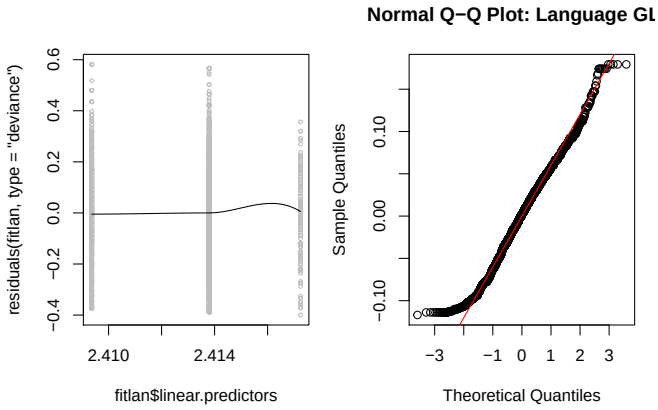
We further test the null hypothesis through a generalized linear model (glm), log-linear regression, as this might provide a different conclusion than the ANOVA test. The log-linear model was chosen because it would return a ratio relative difference in the mean Salary for the groups compared to the 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.

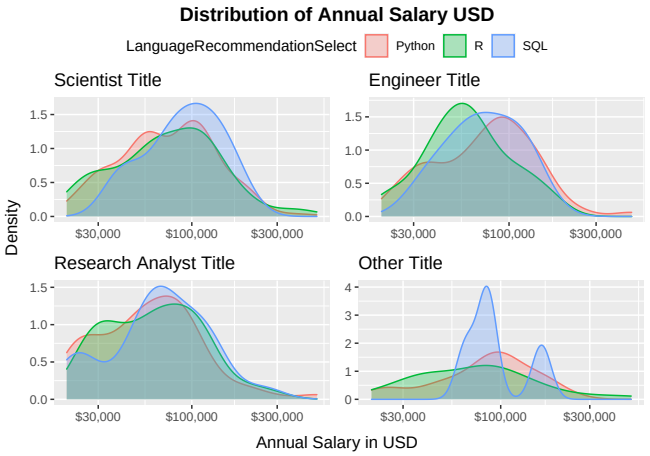
From the Poisson glm for Job Titles, we got that when compared to the mean salary of the *Scientist* group, *Engineers* are ~3% more likely to have lower mean Salary, *Researchers* are ~16% more likely to have lower mean Salary, and those in the *Other* group are ~14% more likely to have higher mean Salaries than Scientist.

For the Language Preference glm, we compared the results to the mean Salary of those who recommended *Python*. Those who recommended *R* were ~4% more likely to have a lower mean Salary, and those who recommended *SQL* were ~1% more likely to have lower mean Salary than the Python group.

The results from the Poisson GLMs were interesting, as they provided an different interpretation to our question. However, when checking for the *dispersion parameter*, we saw that both models were overdispered (jobs dispersion: 3.4672133×10^4, language dispersion: 3.5208343×10^4), meaning that the SEs and test statistics obtained were far from valid. Addressing the assumption of the poisson regression, we can see in the figures below that the assumption of proportionality between mean and variance for poisson regression is not true for the models.





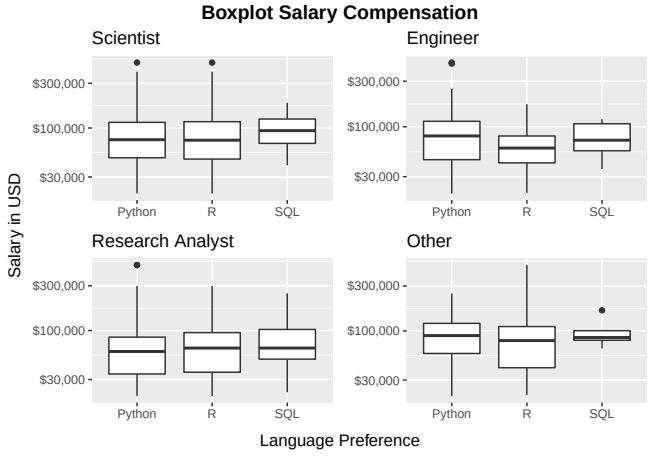


Is the mean Salary within similar Job Titles affected by the choice of Programming Language?

From the Welch t-test, we obtained that for each of the Job Titled groups we no had p-values < 0.0166667, for our Bonferroni correction coefficient. Hence in each of the groups we failed to reject the null hypothesis, since there was no significant evidence for a difference in mean salary for Programming Language recommendation.

For the secon part of the question, given that there was evidence through ANOVA that mean salary between Job Titles was different, we wondered if language recommendation would have an effect in the salary within each Job Title group. In essence, we wanted to test for a difference in mean salary between Scientist who recommend Python vs R vs SQL, and so on for each Job Title.

When visually inspecting each groups variance and distributions (seen in figure below), it was clear that the equal variance assumption for ANOVA would be violated. The Welch t-test was then used for the test and a Bonferroni correction was applied for the 3 group pairs: Python v R, Python v SQL, R v SQL.



Pay & Education

For our third question, we decided to look for both the overall presence of differences in salary between education levels, as well as a linear relationship between education level and salary. The original dataset contains several response categories for education level, a few of which we removed for our analyses. We removed these categories for two reasons: they have small sample sizes, and they do not represent groups we were interested in. The resulting set of education levels is:

- Professional degree
- Bachelor's degree
- Master's degree
- Doctoral degree

The smallest sample size among these groups is about 100, which is large enough to overlook the usual requirement for normal distributions in some of the following analyses.

Previous to actually running Welch T Tests or ANOVA to detect a difference between groups, we performed power caclulations on each approach. Specifically, we calculated the power to detect a difference two groups given a maximum difference of $20,000 among the groups overall (with a significance level of 5%).

Power Calculations

For the multiple Welch T Test approach, we used simulation to calculate power. We generated simulated samples for four groups from normal distributions with means

ranging by $20,000, and sample sizes and variances similar to those from the groups in the real dataset. For each iteration of the simulation, a Welch T Test was run on each pair of groups. The resulting p values were compared to a significance level of 5%, reduced by Bonferroni correction. The rejections were averaged for each test and an overall rejection rate was calculated. The resulting power in our simulations was 0.998.

For ANOVA, we used an R function called *power.anova.test* to generate the power, based on the same parameters as above except that the smallest sample size was used as the size of each group. The resulting power was 0.6928861.

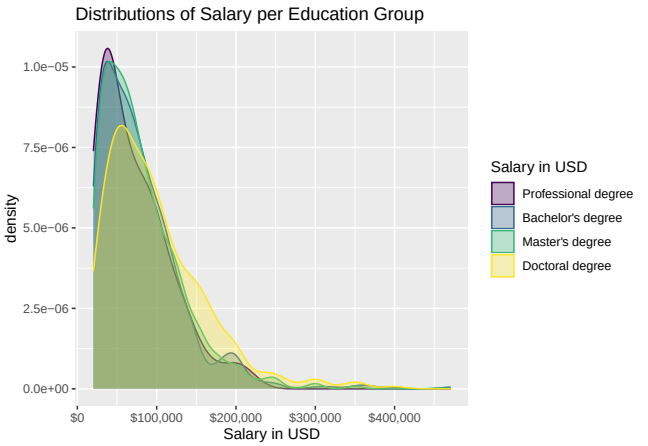
Testing For Difference

On to the actual tests, we first ran the multiple Welch T Tests, one for each pairing of education levels. The resulting p values are shown below. Even with Bonferroni correction, there are three tests whose results indicate a significant difference between groups. Each test between PhD's and other groups was significant, and the rest of the tests were not. Overall, this indicates that there is significant difference in salary between some educational groups, but it further seems to suggest that the only group which is different from the rest is that of PhD's.

	Test	p_value
1	Prof v BS	1.731460e−01
2	BS v MS	5.565941e−01
3	BS v PhD	8.505478e−11
4	MS v PhD	1.806050e−12
5	MS v Prof	8.689002e−02
6	Prof v PhD	1.156592e−06

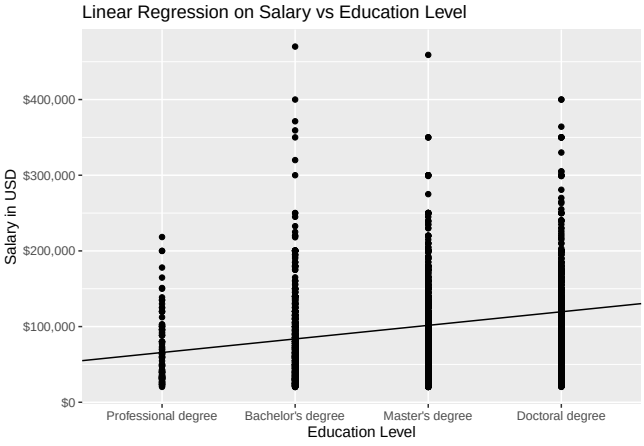
The results of ANOVA also indicate the presence of a difference between groups, with a highly significant p value: $8.6691497 \times 10^{-16}$. There is a caveat to our use of ANOVA which is detailed in the "Discussion" section following this.

The visualization below of the distributions for each group agrees with the conclusions of our tests for differences; there is a difference among groups, and PhD's exhibit the most difference. The conclusions that can be drawn from ANOVA and Welch T Tests, however, are limited to the presence of difference among some groups, and in the case of ordinal data such as education level, there is room to test for a linear relationship.



Linear Regression

We wanted to test for a linear relationship between education level and salary. In order to perform a regression, the education levels were considered to be ordered as follows: Professional, Bachelor's, Master's, and Doctoral. The resulting p value of 9.7238483×10^{-6} is evidence for a linear relationship.



The visualization above graphs the resulting linear relation along with the grouped data. This model's coefficient is 1.793936×10^4, leading to our conclusion that pay is higher by almost $18,000 for each level of increase in education.

Limits of Analyses: The first limitation we noticed was that due to the lack of entries, large amount of zeros, high outliers, and range of low values, we ended up having to significantly reduce the dataset. Since each question only required a few features, all irrelevant features were removed which resulted in the even more data reduction. One major limitation of the analyses presented here is that we report the results of ANOVA tests despite the fact that the groupings of data for each question did not have exactly equal variances. Though this can be a violation of the assumptions of ANOVA, we decided that the differences between variances are not very large relative to the magnitude of the variances themselves. Since there is not a simple rule for how different variances must

be to invalidate ANOVA, we opted to report the results, specifically alongside the results of Welch T Tests, which are valid with the unequal variances.

For the Poisson regression, we saw that some of the assumptions were violated, meaning the results would be questionable and less reliable. The analysis for this part could be addressed by conducting a robust calculation for the SE. Otherwise, the results would be a gross estimate for the data.

We aknowledge that the values of 20k and 30k are not perfect splits for the data into commission and salary features but this type of split was not provided to us. We consider this to be the biggest limitation of the data set since it doesn't make the pay distinction for us. Going forward these values can be changed and varied to find out more information from the dataset or to find what the best value is to split at. If the distinctions were made in the data, we could come up with more accurate answers to our questions.

~~~ 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)
~~~