# Report

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

## Title NUMBER UR PAGES PDF

### Abstract

### Introduction

aims + questions, background material/research briefly For our project we decided to go with the theme of Data Science and Pay. As incoming Data Scientists, we know this career is in high demand but we want to know what features affect the pay of current Data Scientists. To help us gain insight to the relation we found a dataset from Kaggle listed here: https://www.kaggle.com/ikleiman/data-scientists-salaries-around-the-world/data. See the next section for our description of the data. After some exploratory analysis we came up with three questions.

Question 1: Does typical compensation for Data Science professionals differ among densely populated and sparsely populated areas? ('Dense' definition >500 per square mile; 'sparse' definition <500 per square mile). Define density by the number of people that live in an area per square mile.

Question 2:

After analysis of the data during the initial exploration, we noticed that the data should be split into two initial groups (people who are working from commission and those on salary). This split was made because the data contained a large amount of answers that were less than an assumed low salary would be (between 20-40k+). There were also data points that were obvious `outliers` (people who earn 1,000,000 or more) the maximum salary was 28,000,000,000 so we decided to cap those values because we conluded that nobody is making more than 500,000 salary. Another thing that should be noted is that we are only considering people that are employed. The removal of all the participants that reported that earn $0 did not drastically change our results or our final conclusions. ## Data Set Description We had two datasets, the first and primary dataset was from a kaggle project as mentioned above and the other from https://population.un.org/wpp/Download/Standard/Population/. The first dataset is a collection of answers from a global survey given to data scientists and people who have data science related careers. Here is a statement about the dataset "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 project content included schema.csv: a CSV file with 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 does correspond 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 responses in the "Asked" column of the schema.csv file. For the purposes of this analyis the dataset we used was the multipleChoiceReponses.csv.

The second was introduced to answer the first question of density. 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 analysis. source, collection methods, study design (randomized exp, obs, survey, etc)

## Statistical Methods

Methods: 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. 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. After some exploratoy analysis of the data we found that the variances we not quite equal which lead us to conduct a Welch T-Test for each of the questions as well.

For question 1, we wanted to use ANOVA to test possible differences in multiple groups (groups being high and low density) and Linear Regression to test if there is a linear relationship between compensation and the density of groups.

Assumptions: 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.

For question 2,

For question 3,

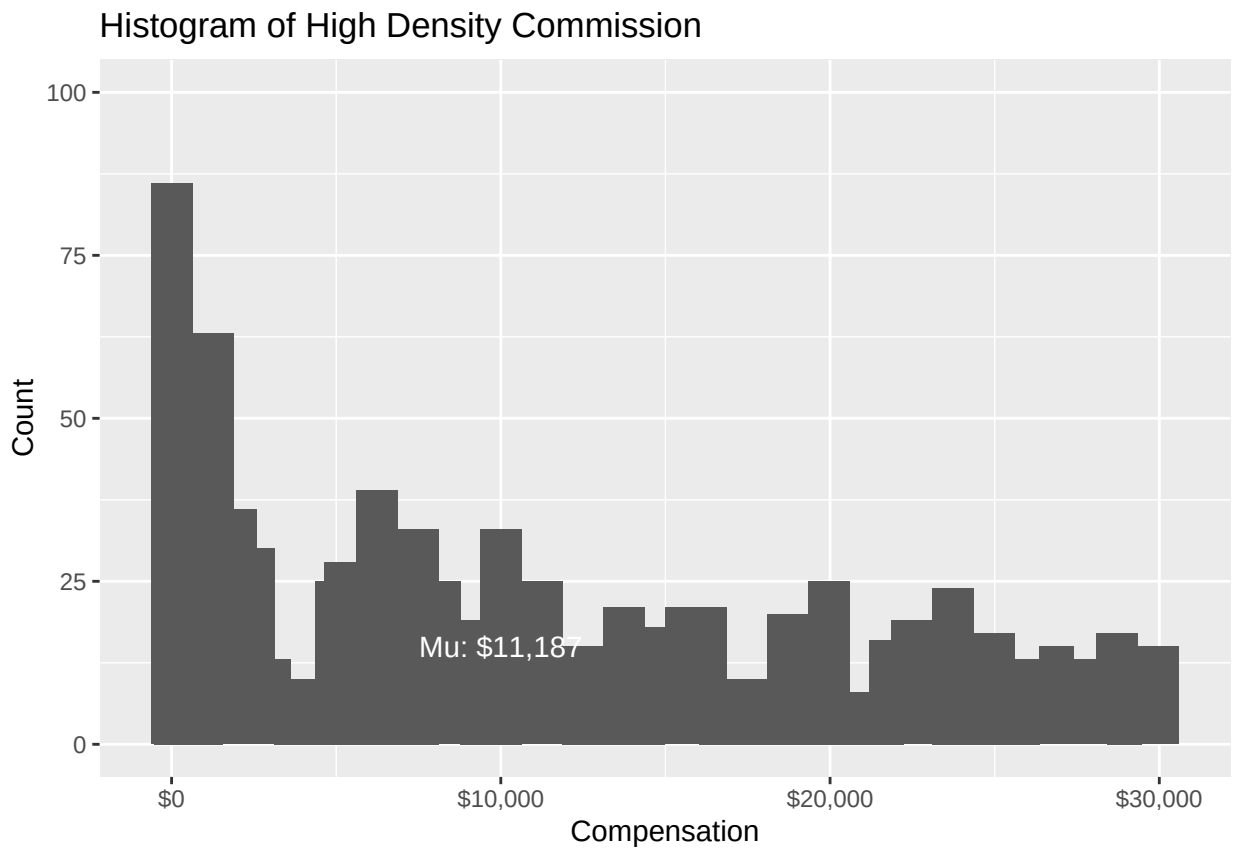
Explanation for why methods appropriate

## Results

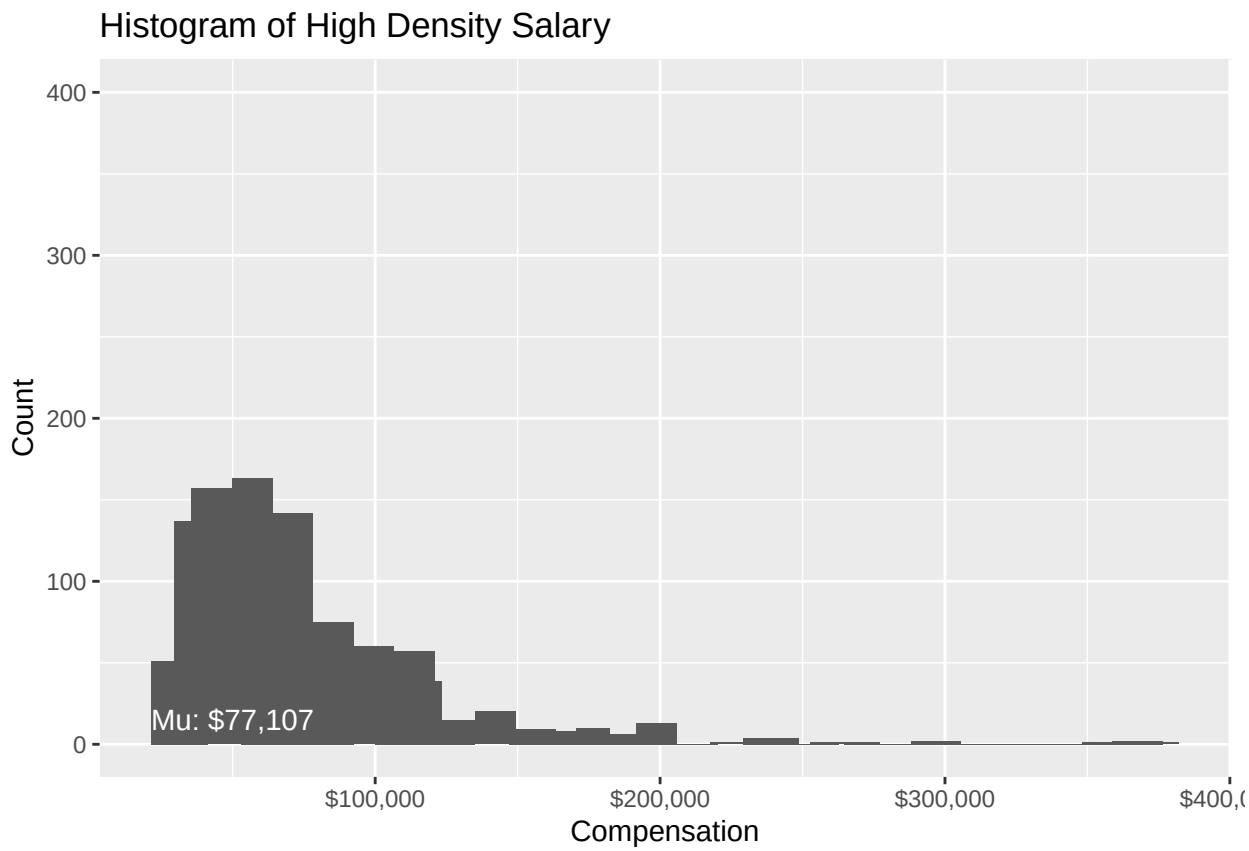
graphs, tables, descriptive info about data and results of analysis ## Results from Question 1

Here are the distributions for commission and salary for high denstiy and low density.

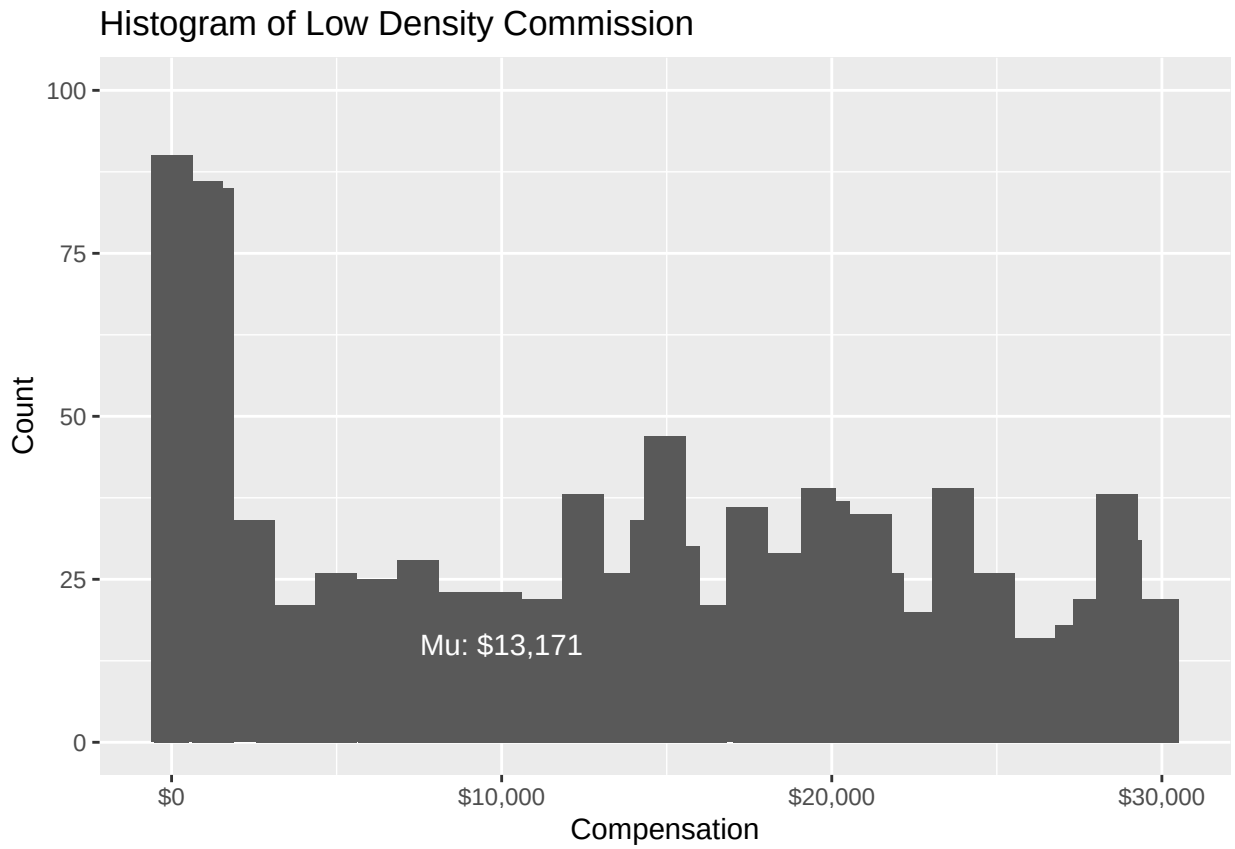
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



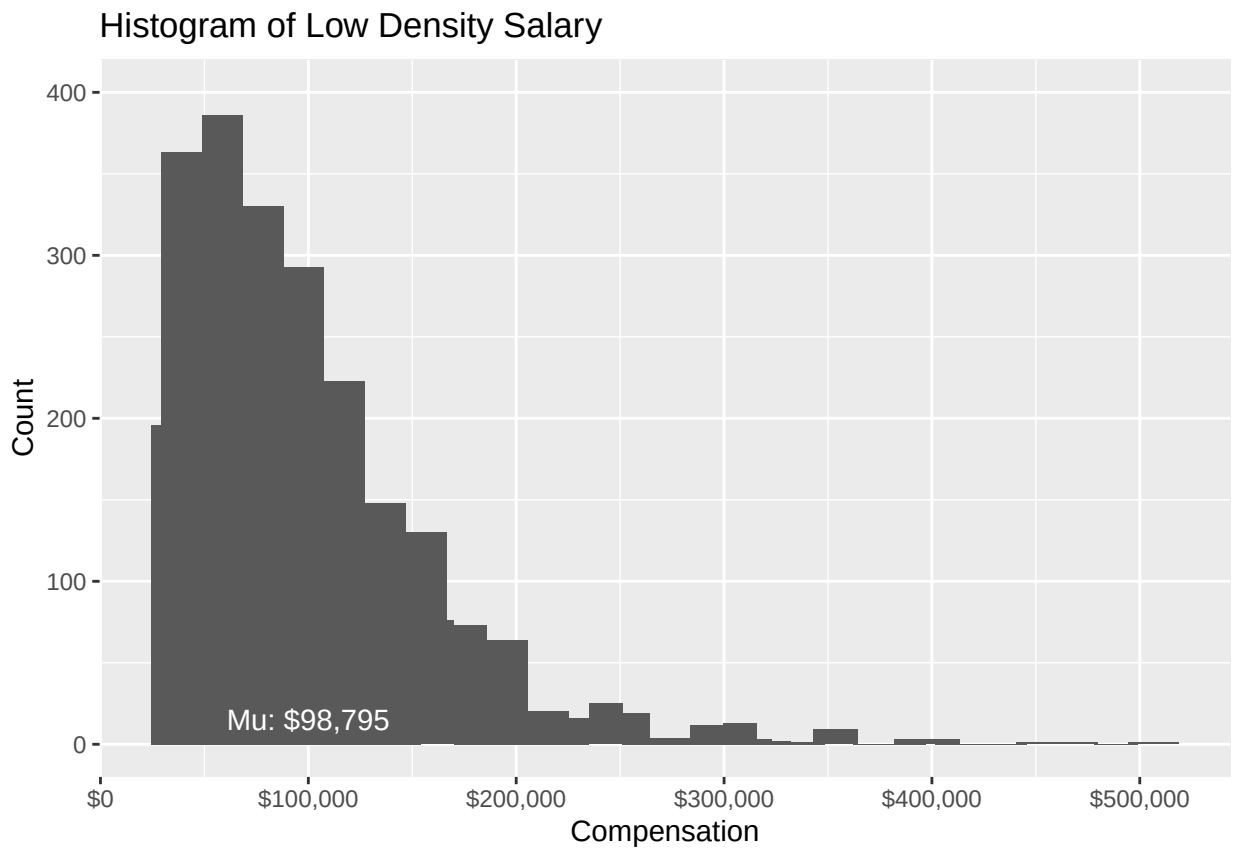

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

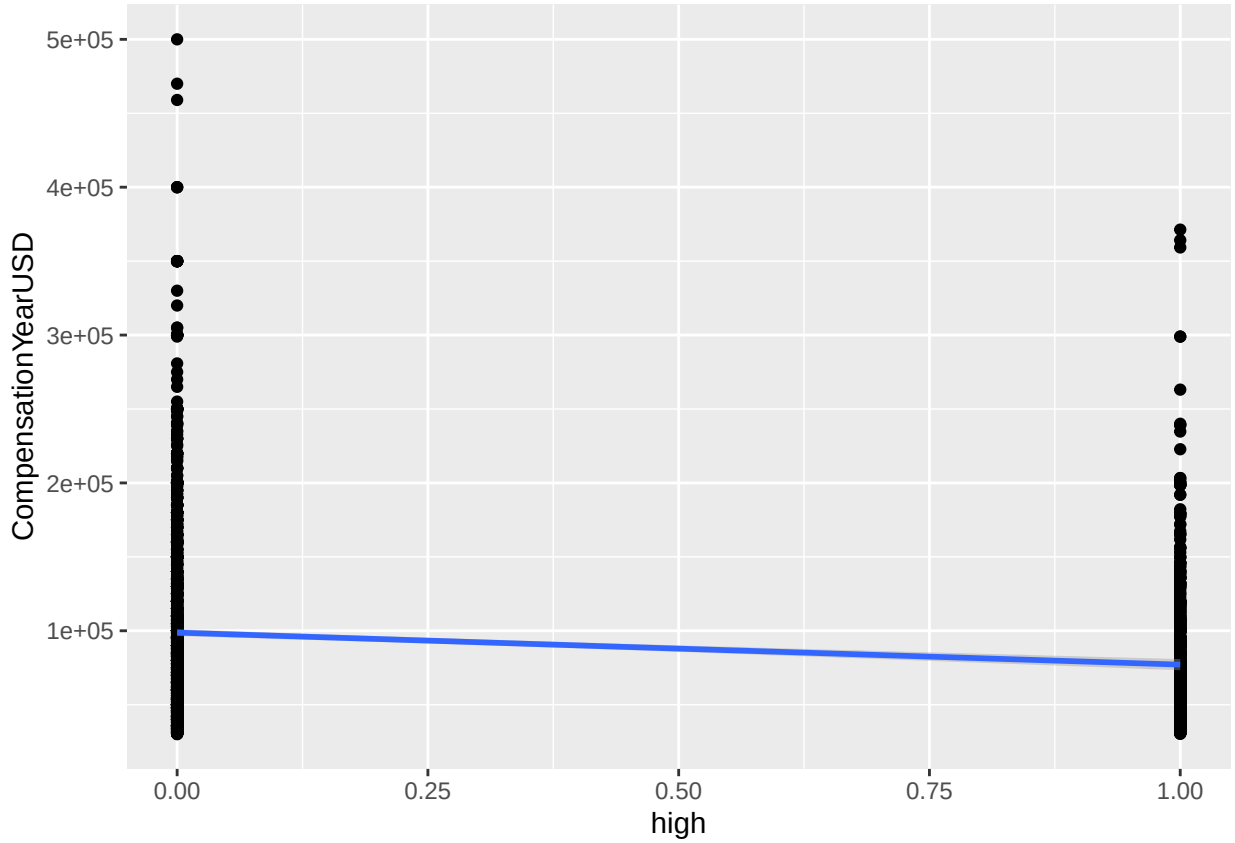



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

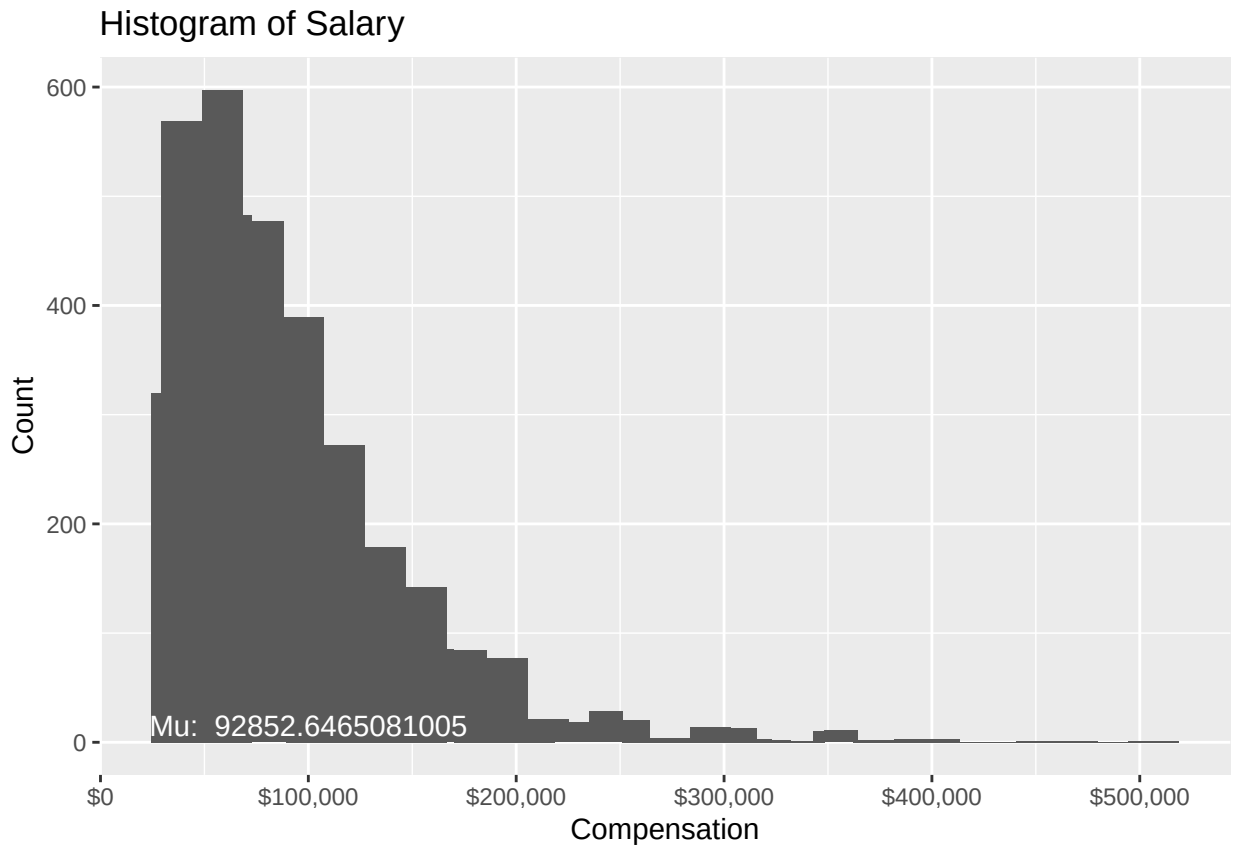



From Linear Regression, there is a slight negative linear relationship between density and compensation (for salary and commission). As density goes up on average Salary goes down by $21,652.

```
## `geom_smooth()` using formula 'y ~ x'
```

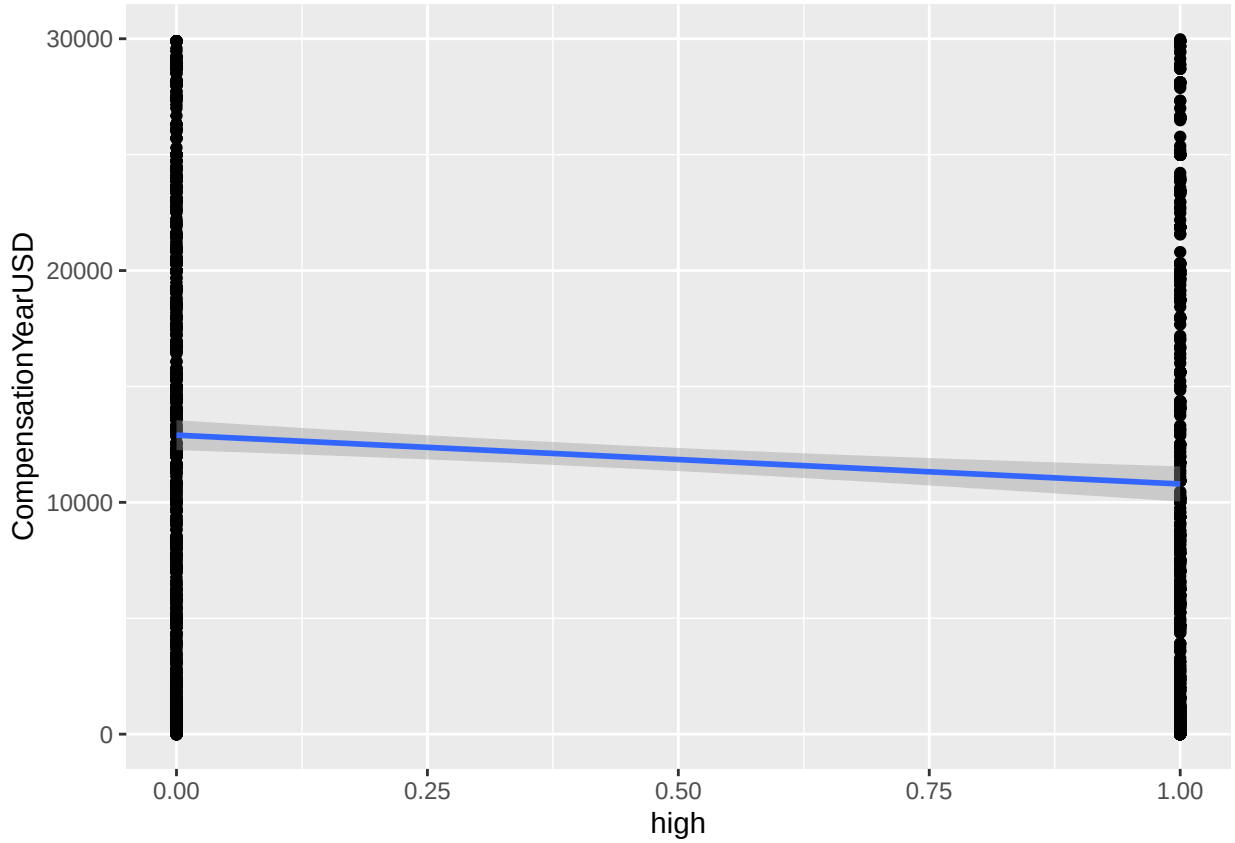



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

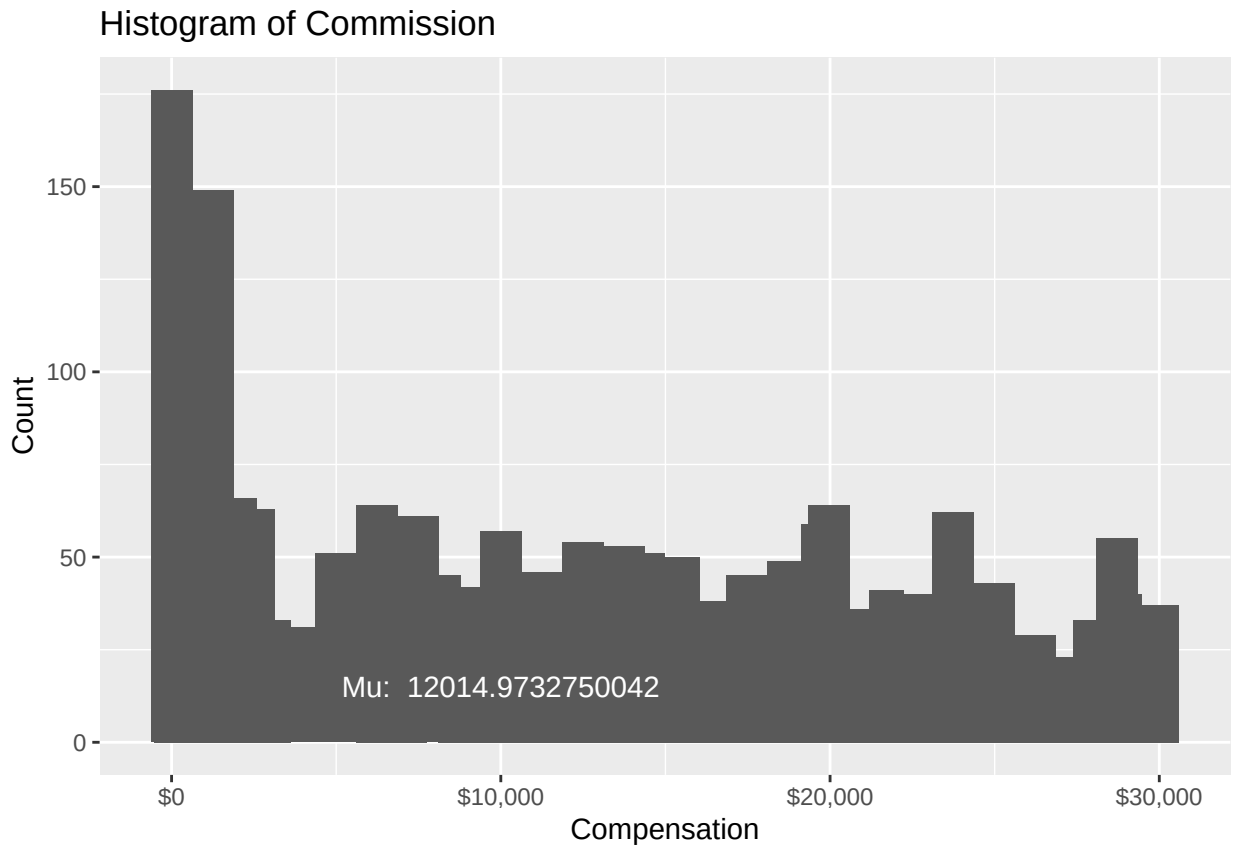



This is the linear regression plot and the distribution of the total commission (including high and low density), similarly as density goes up on average commission goes down by $2,106.

```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




### Results from question 2:

### Results from question 3:

### 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