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. 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 salary would be (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. ## Data Set Description source, collection methods, study design (randomized exp, obs, survey, etc)

Statistical Methods

Methods: Anova, Welch T-test, linear regression 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,

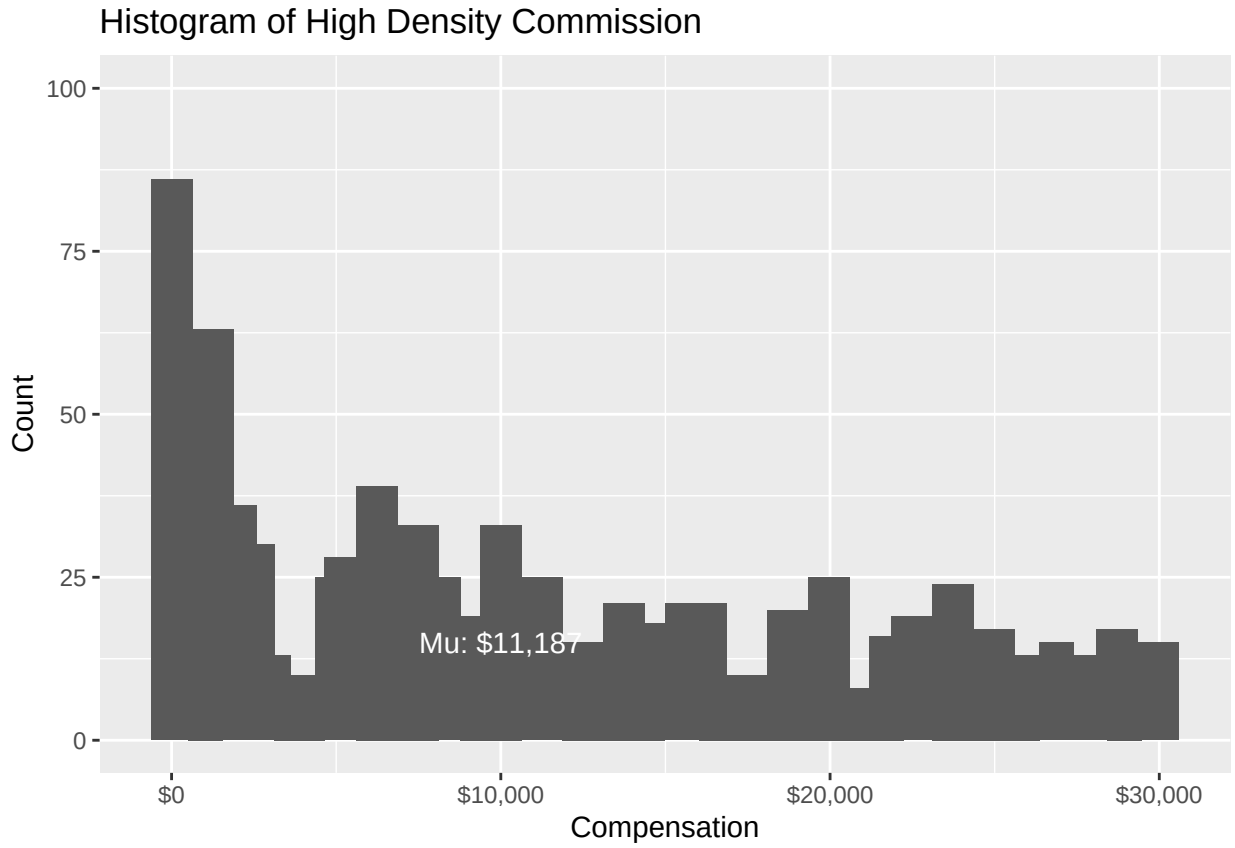
Explanation for why methods appropriate

Results

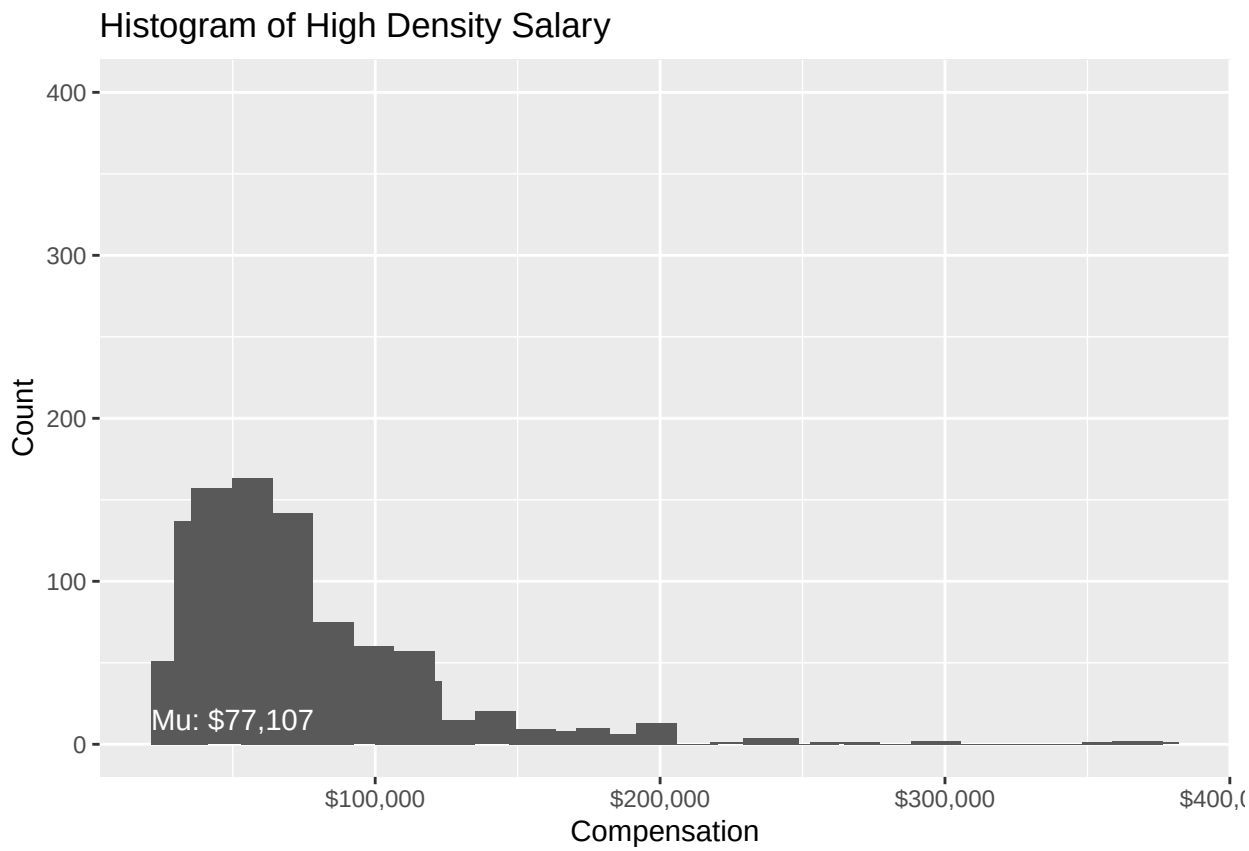
graphs, tables, descriptive info about data and results of analysis ## Results from Question 1

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

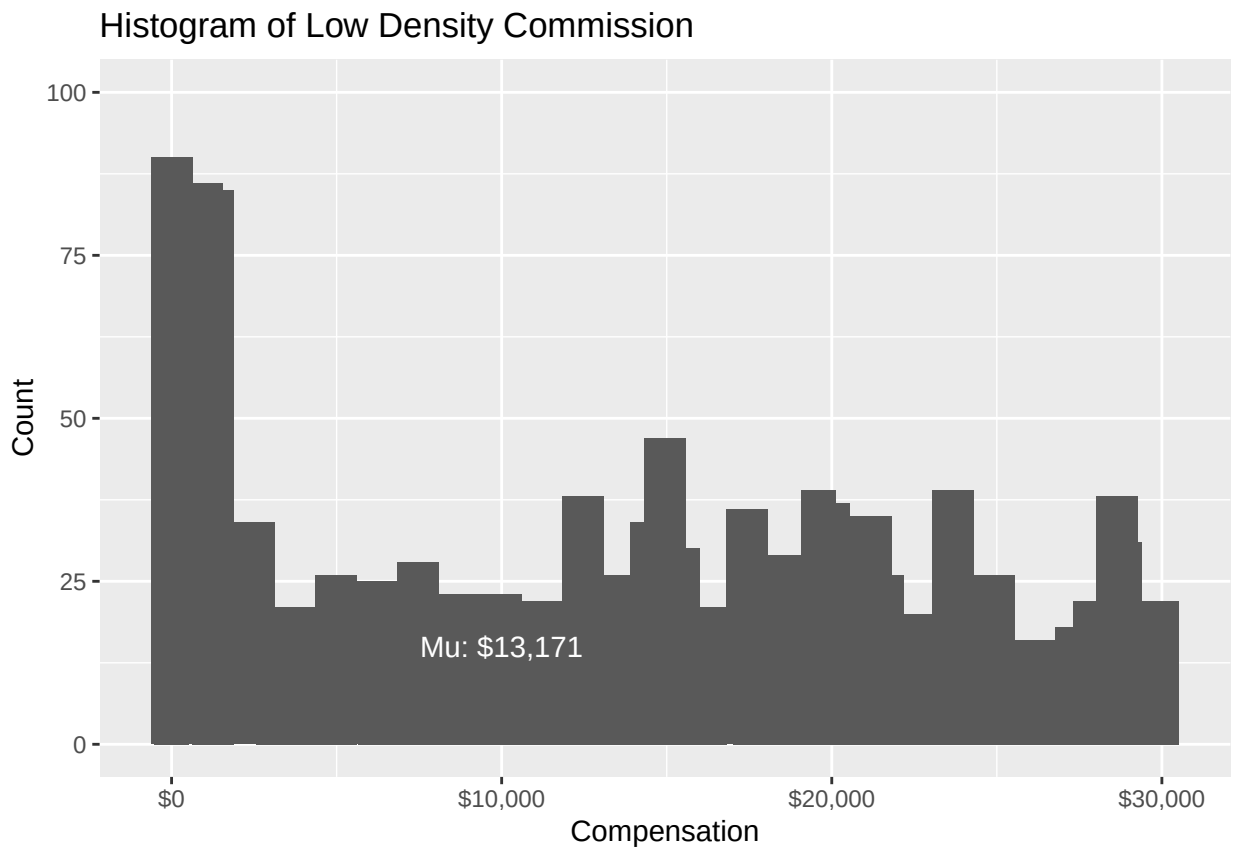
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



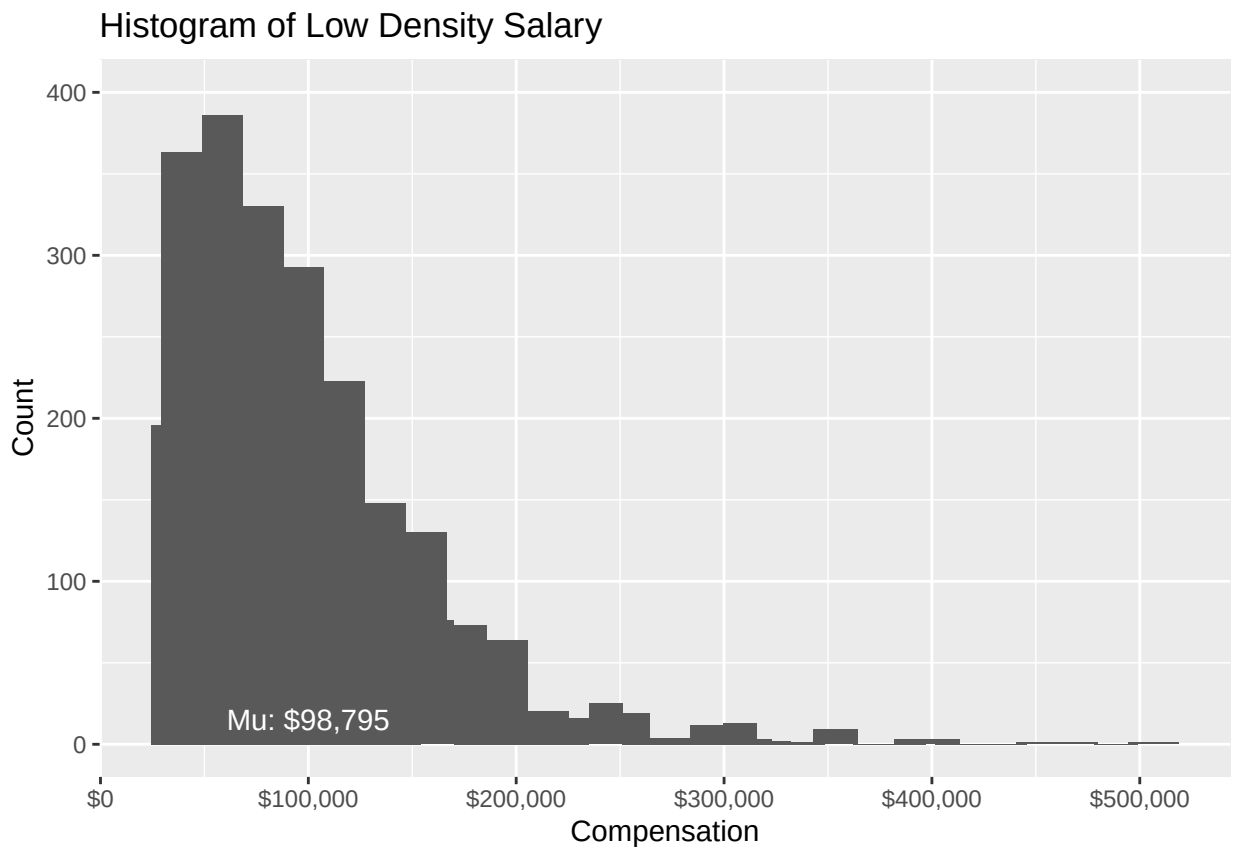

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

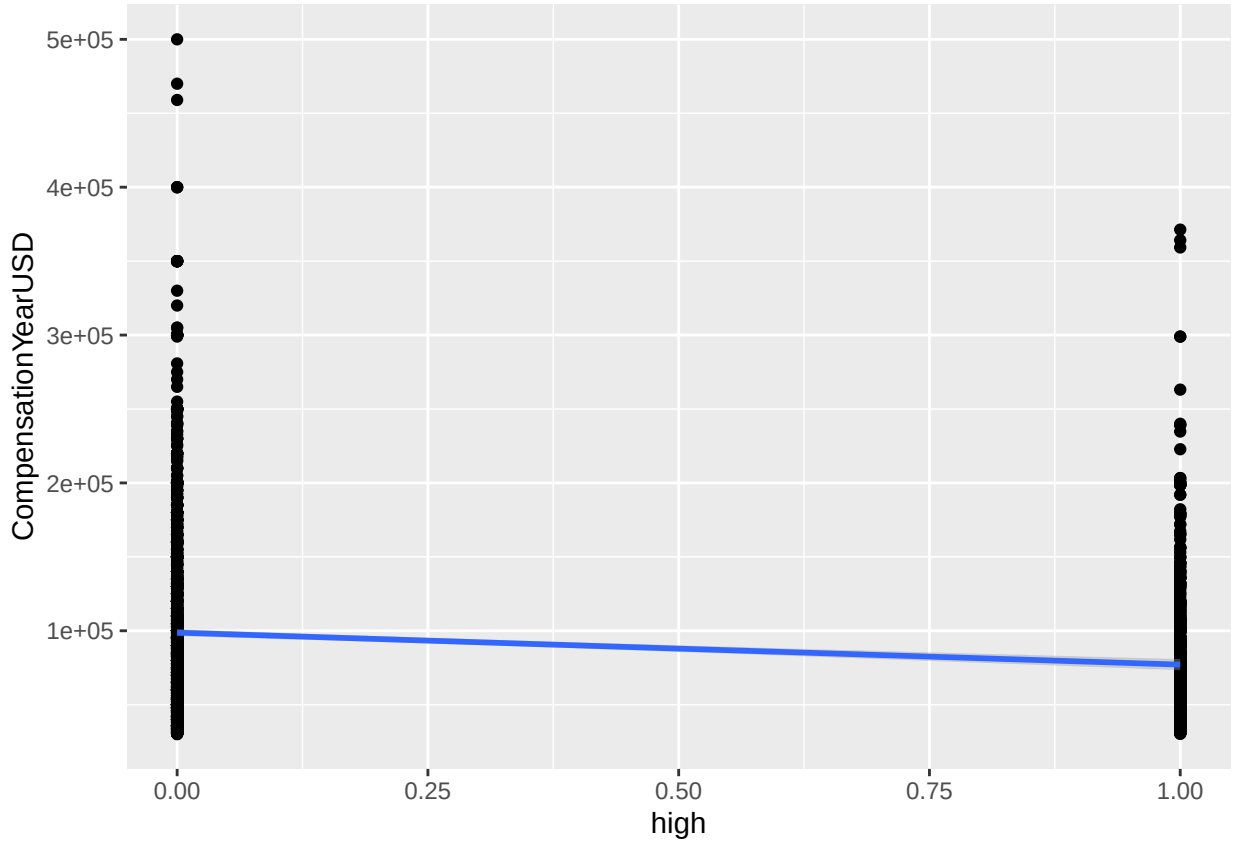



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

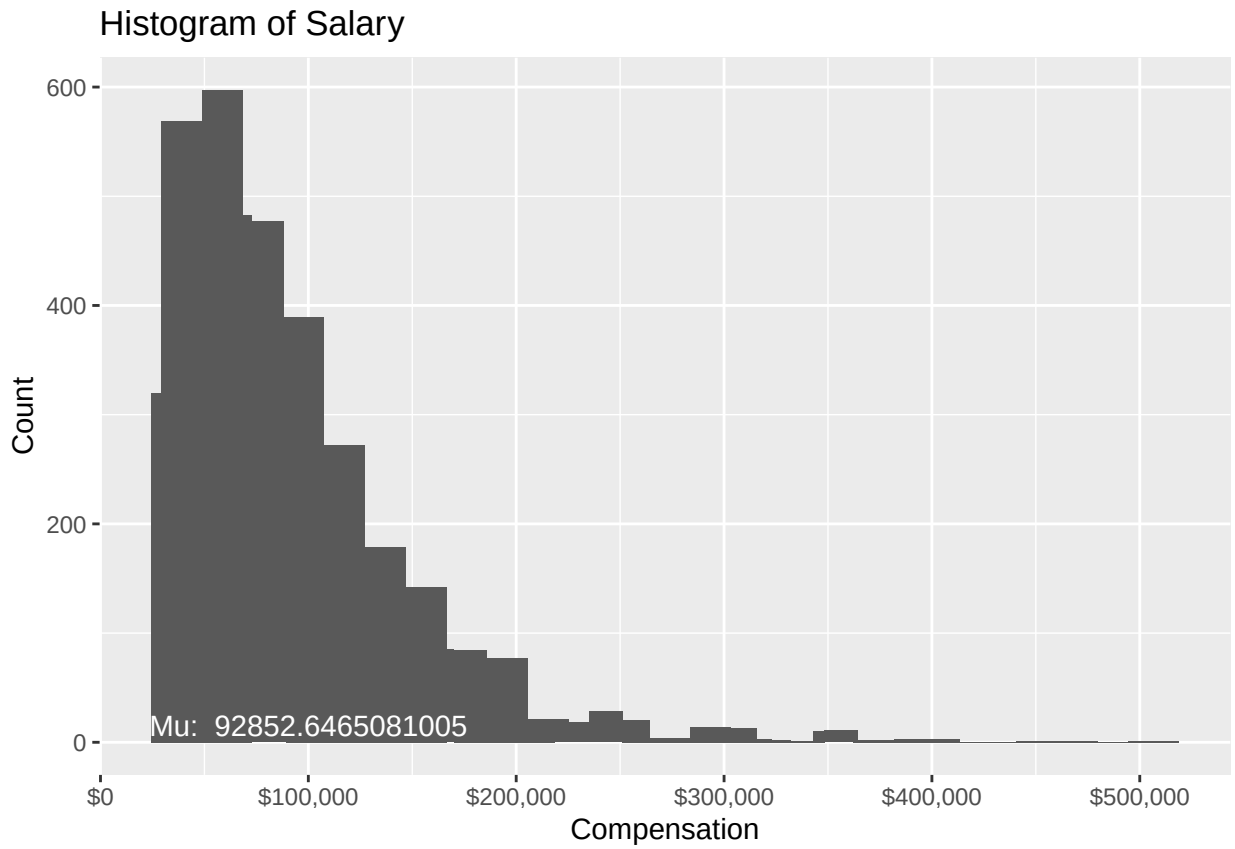


This is the linear regression plot and the distribution of the total salary (including high and low density)

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

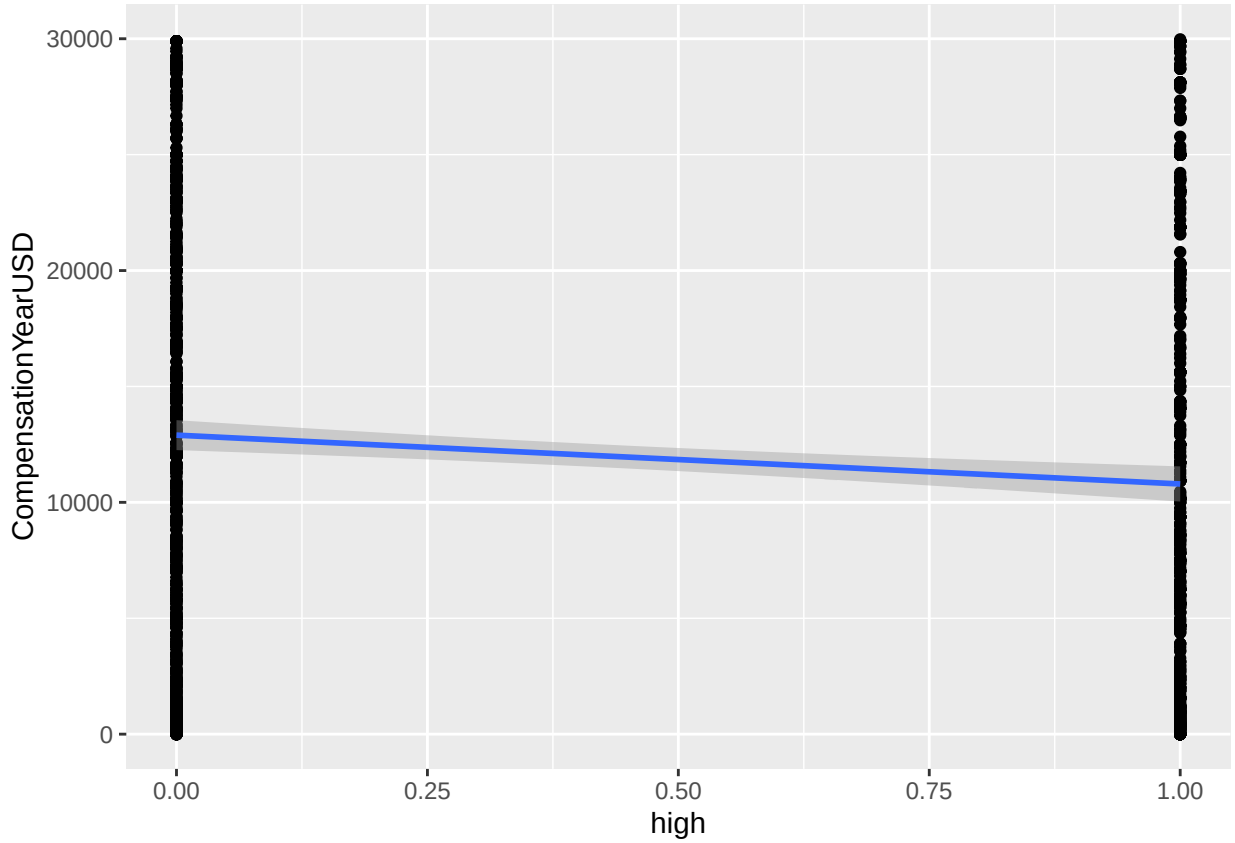



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

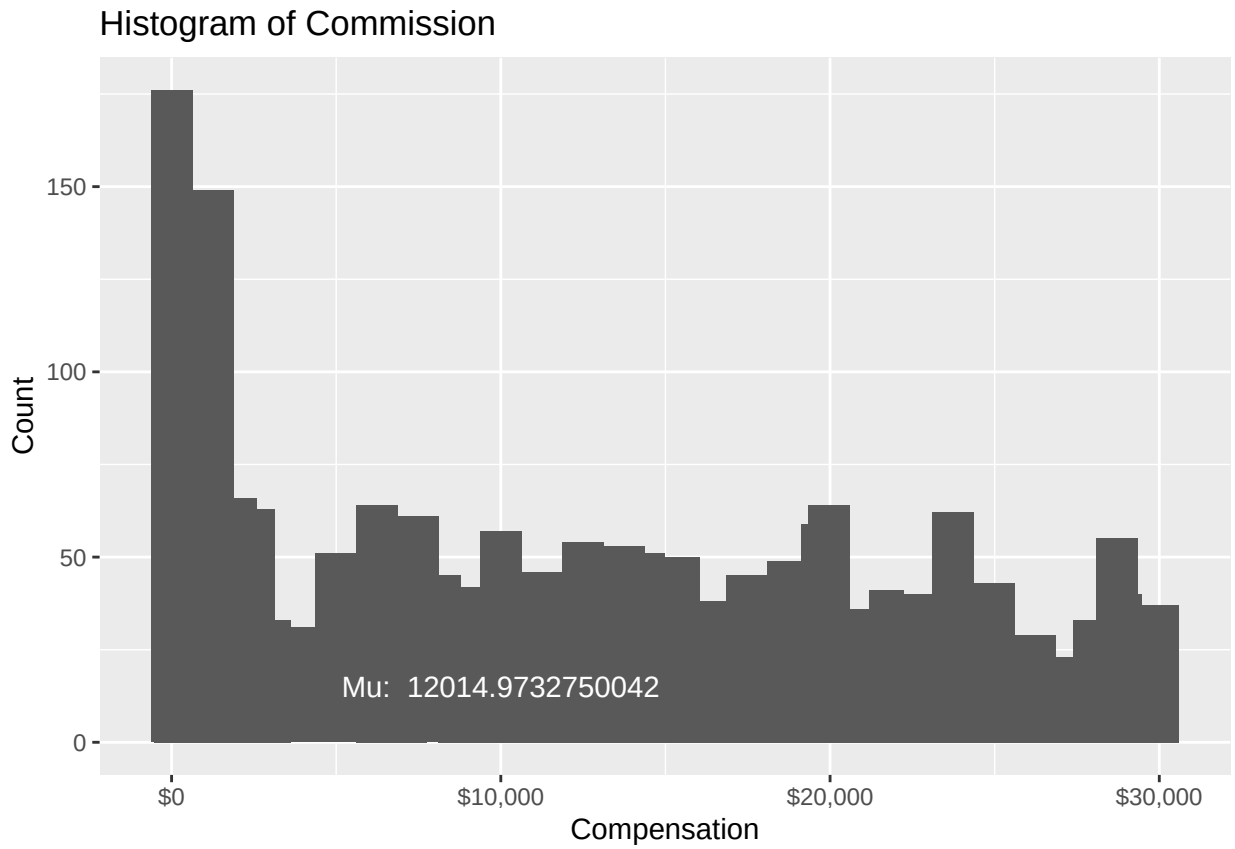


This is the linear regression plot and the distribution of the total commission (including high and low density)

```
## `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: 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 tests anyway. The p-value reported from the anova tests are <2e-16 for salary and .00011 for commission. From Linear Regression, there is a slight negative linear relationship between density and compensation (for salary and commission). As density goes up Salary goes down by $21,652, and similarly as density goes up commission goes down by $2,106. 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.

Question 2:

Question 3:

limits of analyses: Question 1: Unequal variances.

Question 2:

Question 3:

limits of data too: Question 1:

Question 2:

Question 3:

References

brief

Appendices

more technical aspects of analyses, any other tidbits