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

Data Set Description

source, collection methods, study design (randomized exp, obs, survey, etc)

Statistical Methods

methods, discussion of assumptions, explanation for why methods appropriate Anova, Welch T-test, linear regression

Results

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

#{r setup, include=FALSE} #knitr::opts_chunk$set(echo = TRUE) #

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 amount of people that live in an area per square mile for a given country.

Use ANOVA to test possible differences in multiple groups and Linear Regression to test if there is a linear relationship between compensation and the density of groups.

Test level of power data set allows for our research given the n number of responses.

```
## Warning: package 'naniar' was built under R version 3.6.3
```

Interesting fact: We have significantly more datapoints for low density locations than high density locations. We initially thought that this would be a problem.

Also we are not including those people that have shared that they are unemployed in the commision calculations, we found that doing this did not drastically

change our data since the people that reported that they make 0$ was very low. It also made sense to only include those people that are working in the calculation.

Note that there is a high amount of varaince in the data. This occurs because there are people who are reporting that they make one dollar and people that they are

making the upper limit we have capped the data at for commision. Similarly for salary we have people that have reported that they make close to half a million and

people that report that they make the minumum we have set of $30,000.

```
# Sample size estimation for question
delta1 <-  low_salary2_mean - high_salary_mean
delta2 <-   high_commision_mean - low_commision2_mean
salary_n = ((low_salary2_var + high_salary_var)*  ((qnorm(1- 0.025) + qnorm(.9))**2)   )/(delta1**2)
commision_n = ((low_commision2_var + high_commision_var)*  ((qnorm(1- 0.025) + qnorm(.9))**2)   )/(delta
```

Note that we only need a sample size of 138 for commision to achieve a power of .9 and a sample size of 290 for salary to achieve the same power. This is why

achieve the power that we have calculated from our experiment.

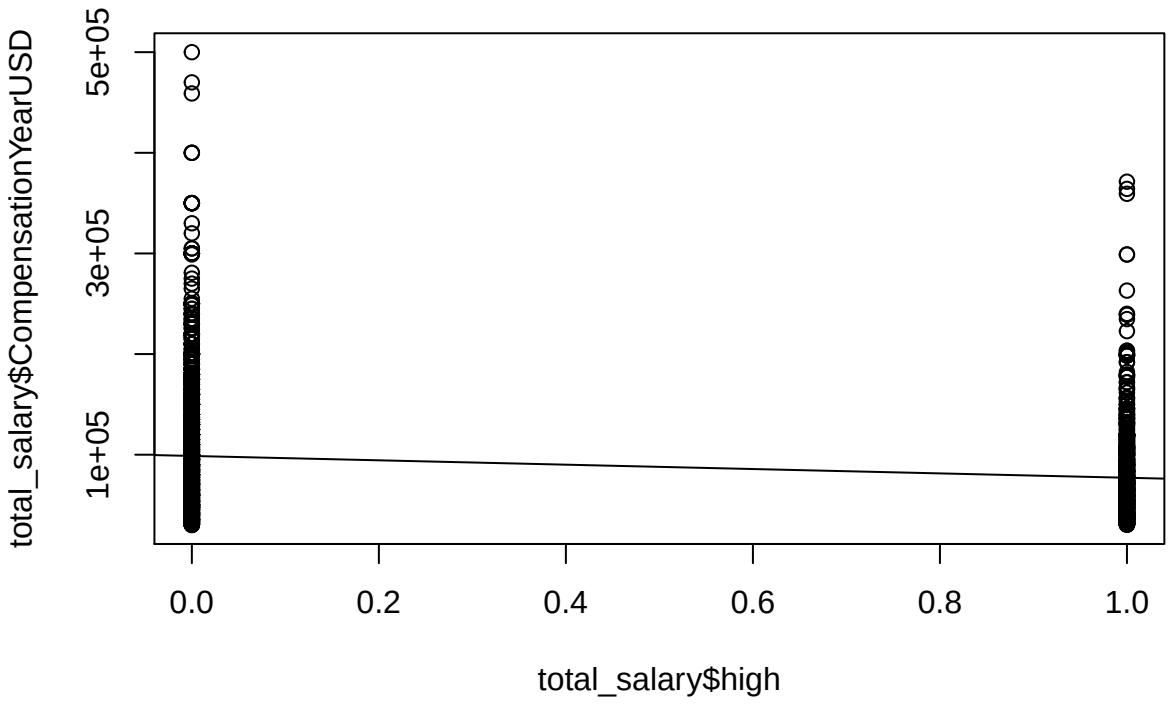
```
# Type 1 error probability
```

```
## [1] "Anova"
```

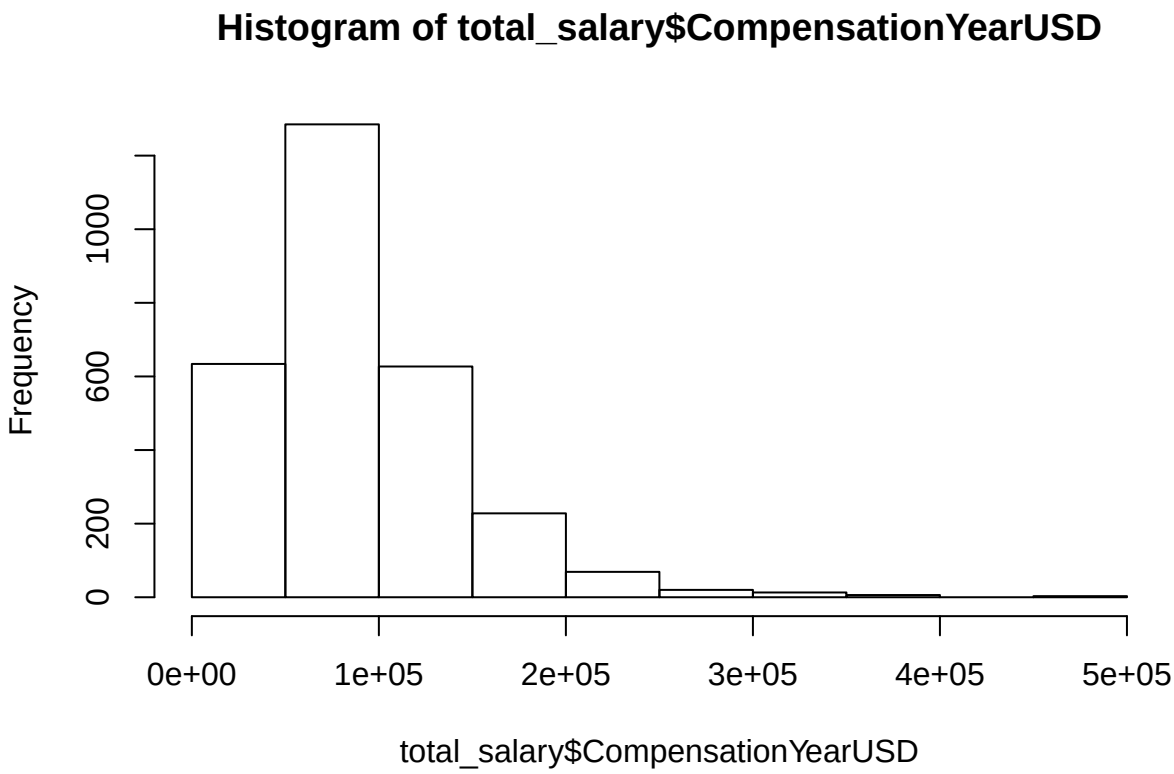
```
## [1] "\nsummary(aov(total_salary$CompensationYearUSD ~ total_salary$high))\nsummary(lm(total_salary$C
```

```
## [1] "\n\nsummary(aov(total_commision$CompensationYearUSD ~ total_commision$high))\nsummary(lm(total_
```

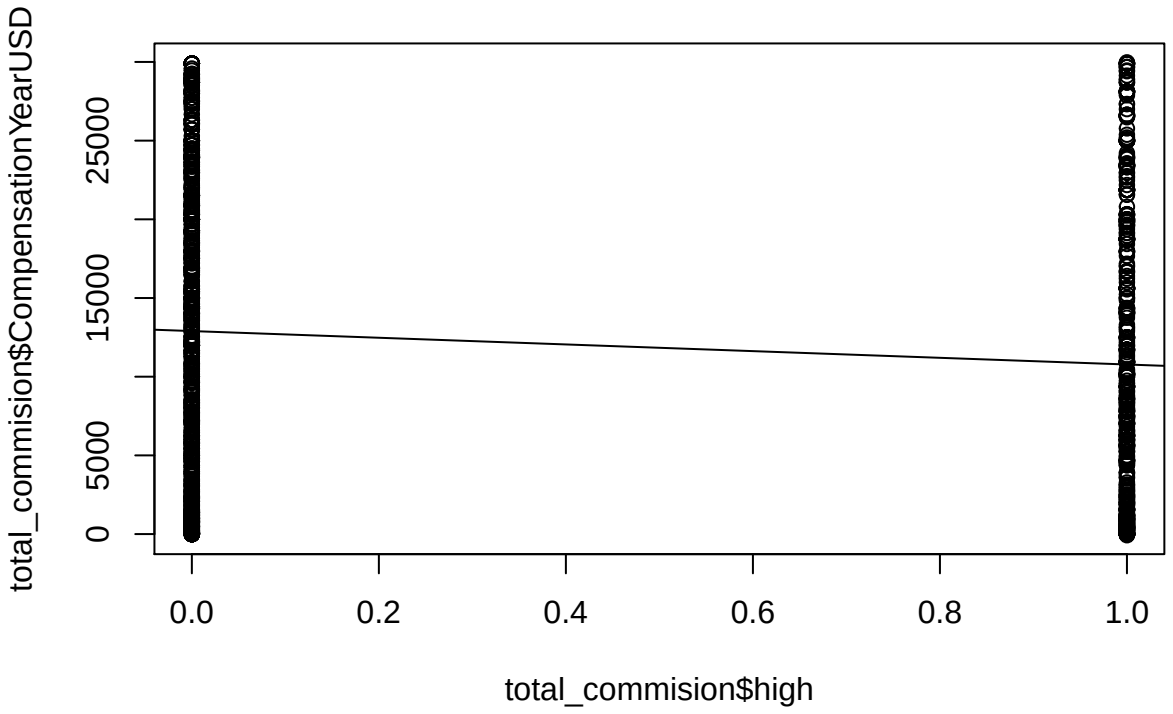
```
# Interesting plots for salary
plot(total_salary$high, total_salary$CompensationYearUSD)
abline(lm(total_salary$CompensationYearUSD ~ total_salary$high))
```



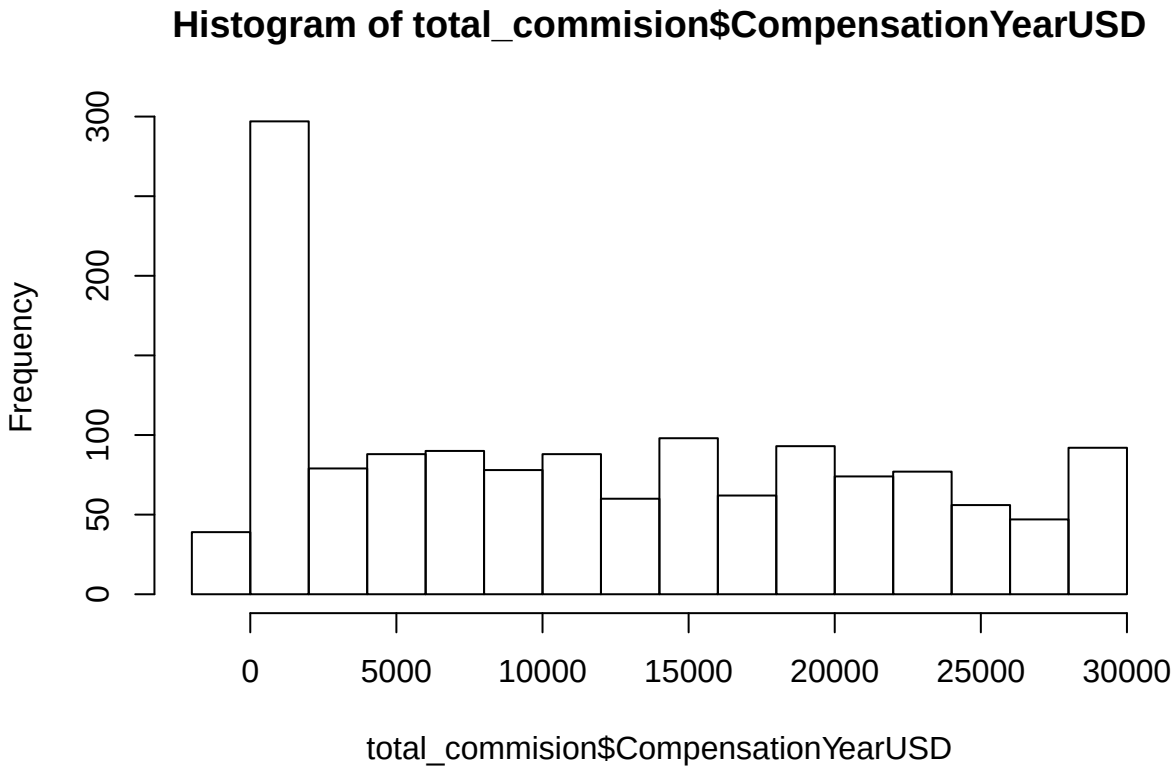

```
hist(total_salary$CompensationYearUSD)
```




```
# Interesting plots for commmision
plot(total_commision$high, total_commision$CompensationYearUSD)
abline(lm(total_commision$CompensationYearUSD ~ total_commision$high))
```




```r
hist(total_commision$CompensationYearUSD)
```



Conclusion from initial statistics, is that we should not use Anova or linear regression because the variances are not the same, and

we know that equal variances are a problem no matter how large the sample size thus we should use Welch.

When looking at the p-value for salary and commision they are 2.592e-12 and 7.519e-11 respectively so there is evidence to reject the null hypothesis

meaning that density has an effect on compensation for salary and commision.

Discussion

summarize results and conclusions limits of analyses limits of data too

References

brief

Appendices

more technical aspects of analyses, any other tidbits