Report

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

Title NUMBER UR PAGES PDF

Abstract

Introduction

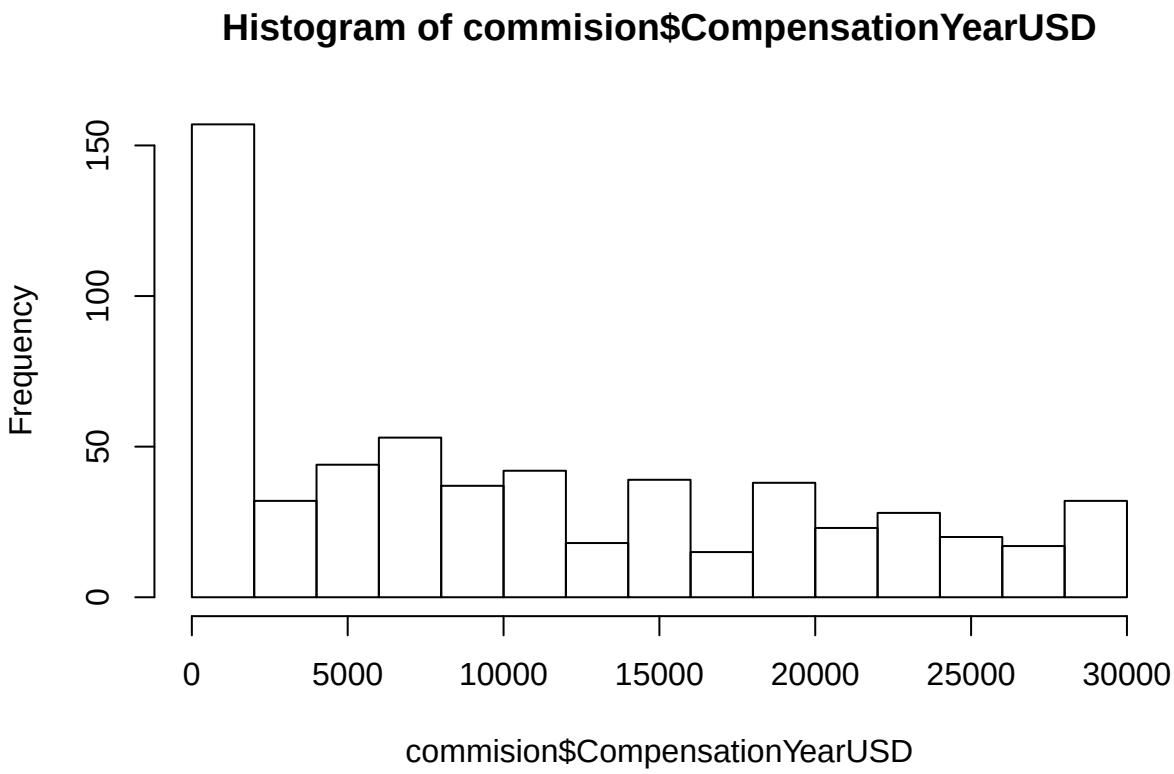
aims + questions, background material/research briefly For our project we decided to go with the theme of Data Science and Pay. To help us answer this question we found a dataset from Kaggle listed here: https://www.kaggle.com/ikleiman/data-scientists-salaries-around-the-world/data. See the next section for our description of the data. After 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)

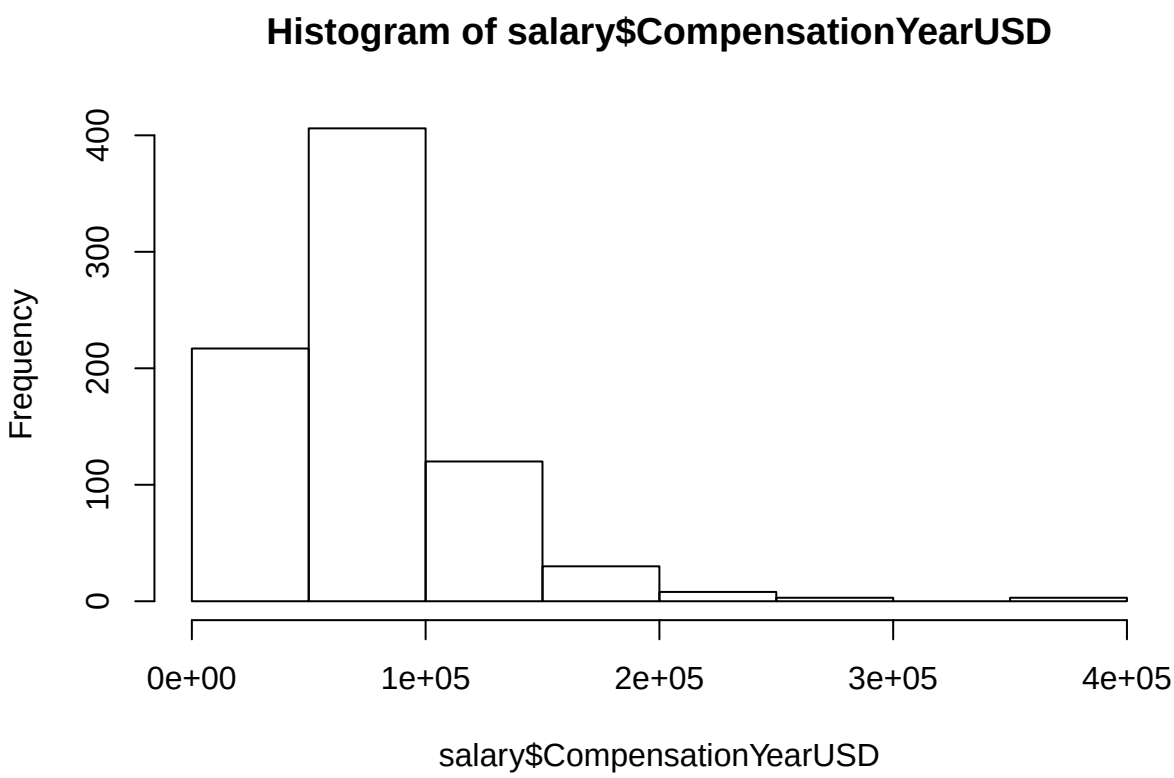
Statistical Methods

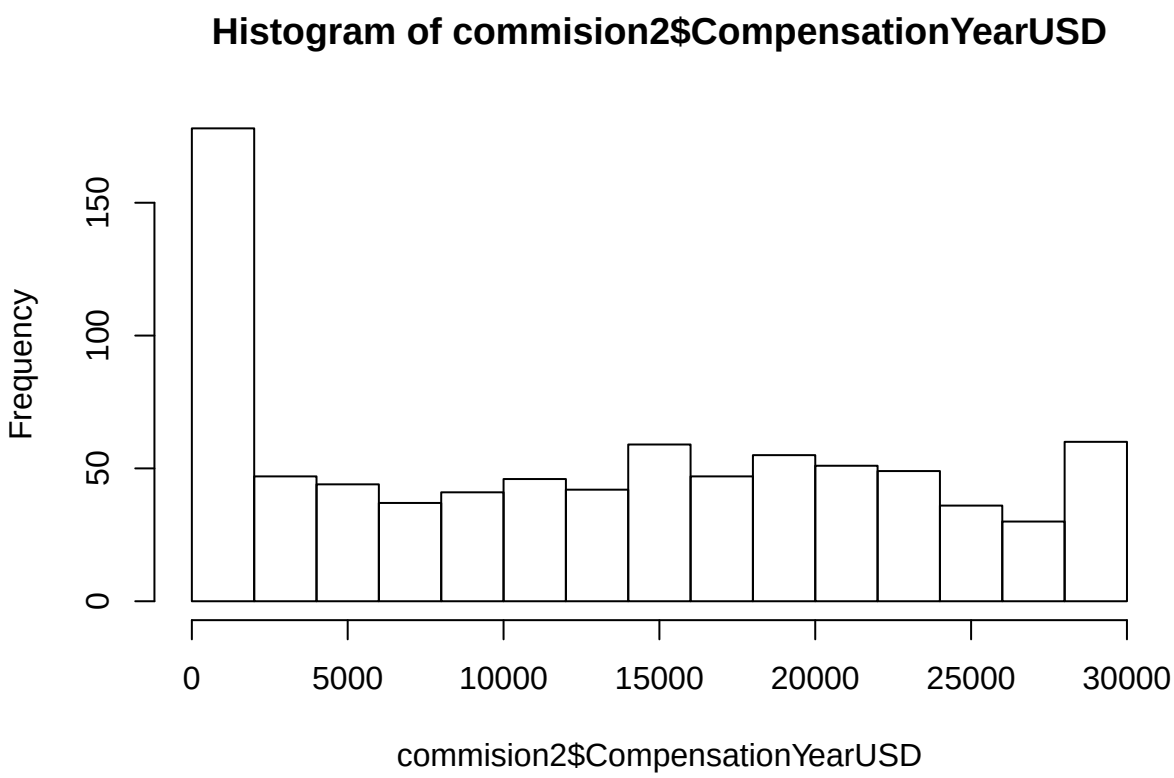
methods, discussion of assumptions, explanation for why methods appropriate Anova, Welch T-test, linear regression

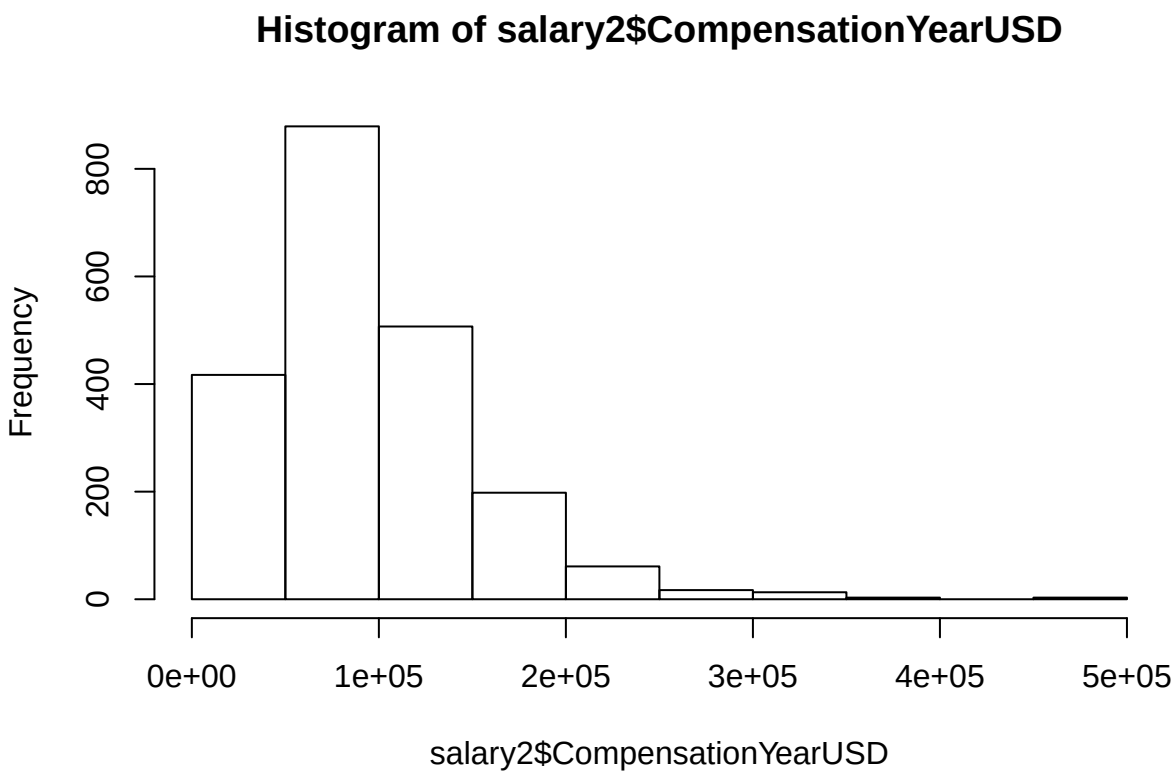
Results

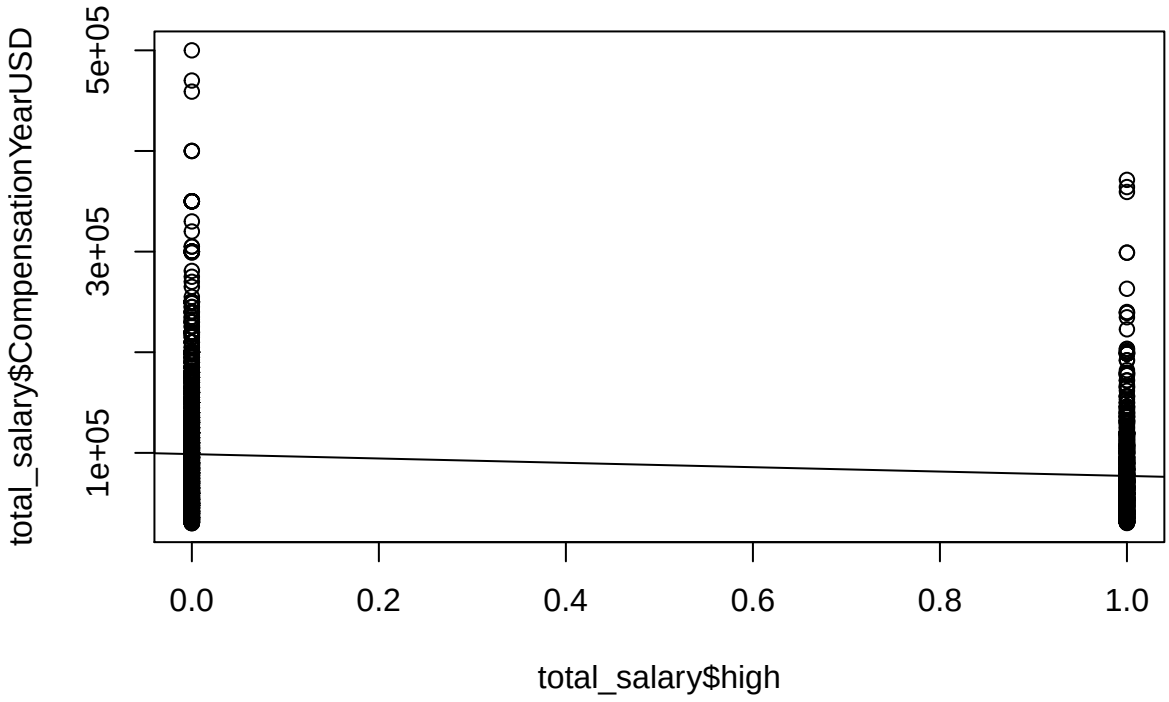
graphs, tables, descriptive info about data and results of analysis # Results from Question 1

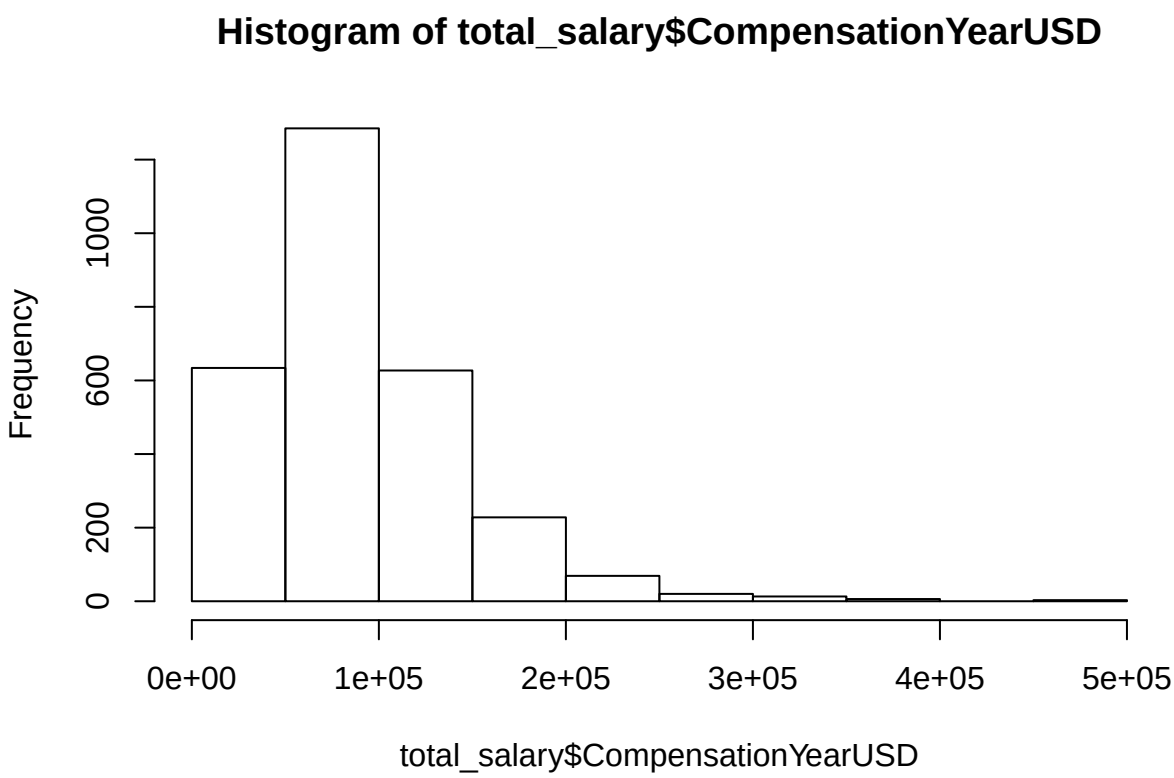

Histogram of commision$CompensationYearUSD
Frequency
0
50
100
150
0
5000
10000
15000
20000
25000
30000
commision$CompensationYearUSD

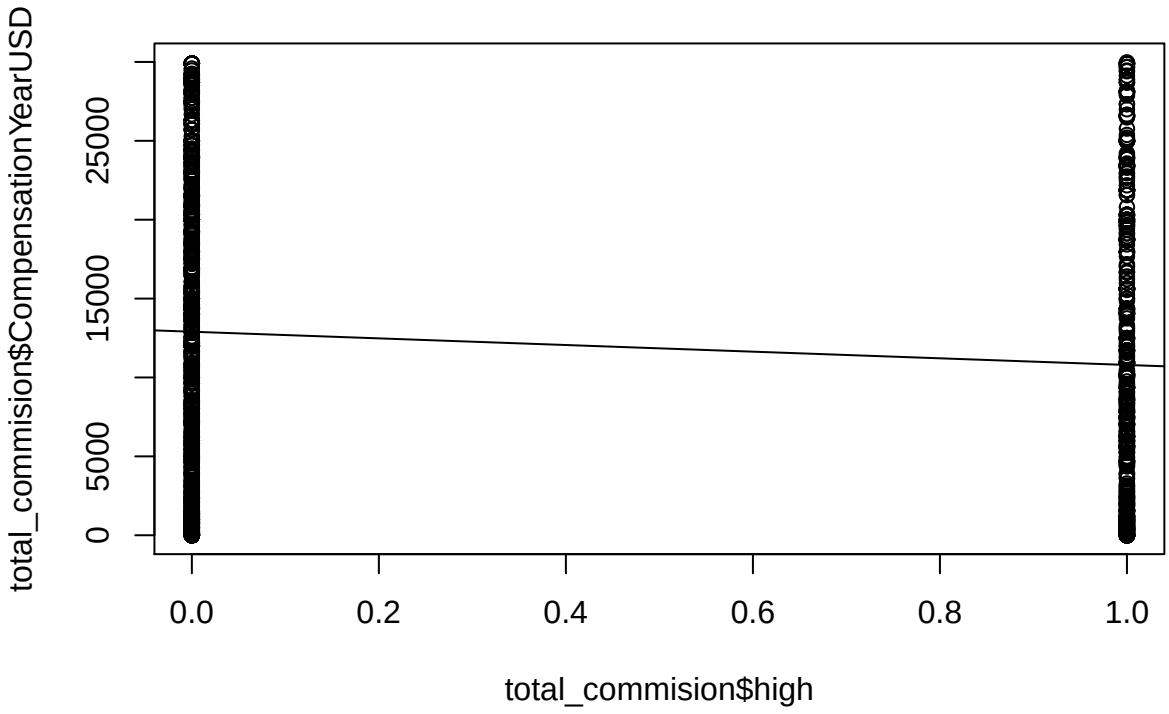

Histogram of salary$CompensationYearUSD
Frequency
0
100
200
300
400
0e+00
1e+05
2e+05
3e+05
4e+05
salary$CompensationYearUSD

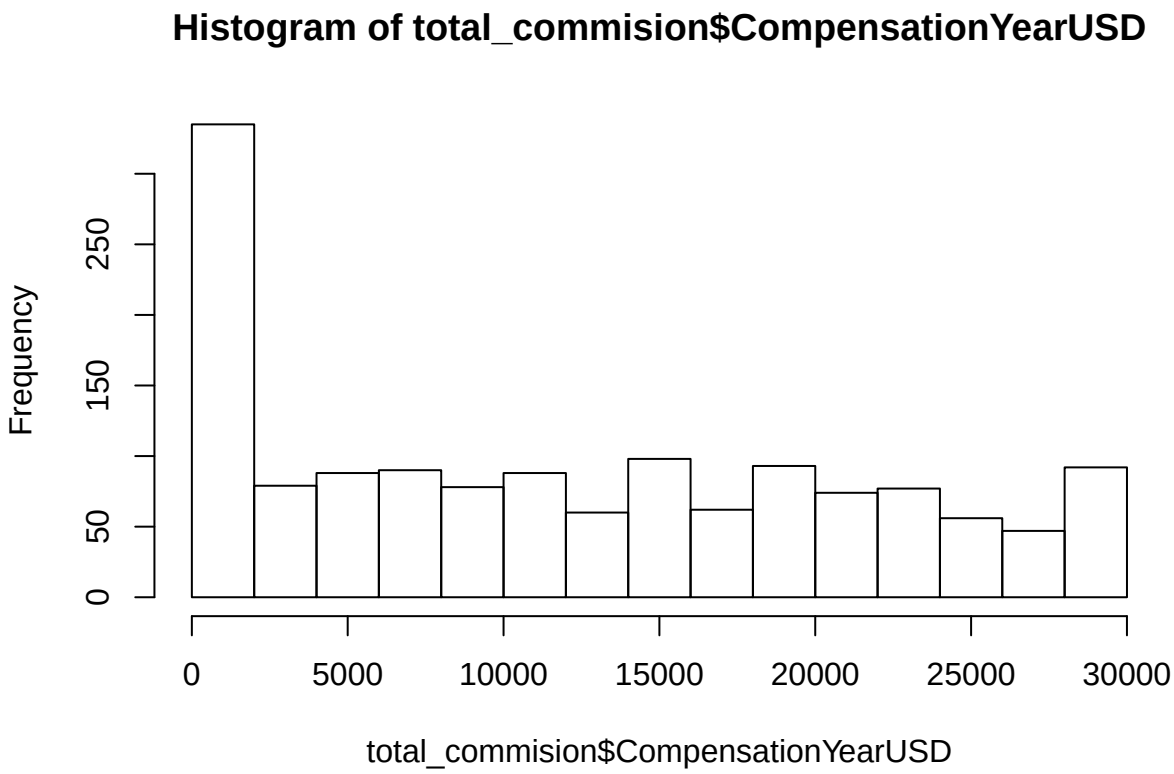

Histogram of commision2$CompensationYearUSD
Frequency
0
50
100
150
0
5000
10000
15000
20000
25000
30000
commision2$CompensationYearUSD


Histogram of salary2$CompensationYearUSD
Frequency
0
200
400
600
800
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
salary2$CompensationYearUSD


total_salary$CompensationYearUSD
1e+05
3e+05
5e+05
0.0
0.2
0.4
0.6
0.8
1.0
total_salary$high


Histogram of total_salary$CompensationYearUSD
Frequency
0
200
600
1000
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
total_salary$CompensationYearUSD


total_commision$CompensationYearUSD
0
5000
15000
25000
0.0
0.2
0.4
0.6
0.8
1.0
total_commision$high



Discussion

summarize results and conclusions limits of analyses limits of data too

References

brief

Appendices

more technical aspects of analyses, any other tidbits