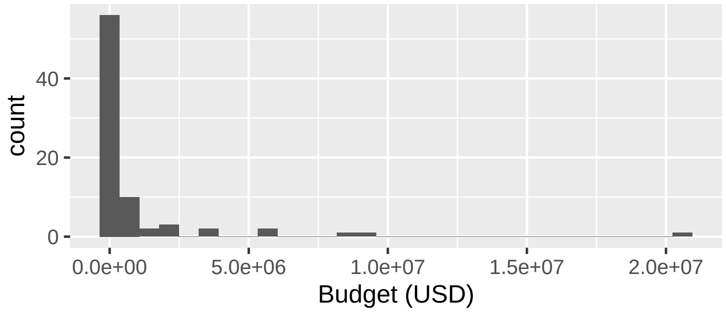
Pair Assignment 2

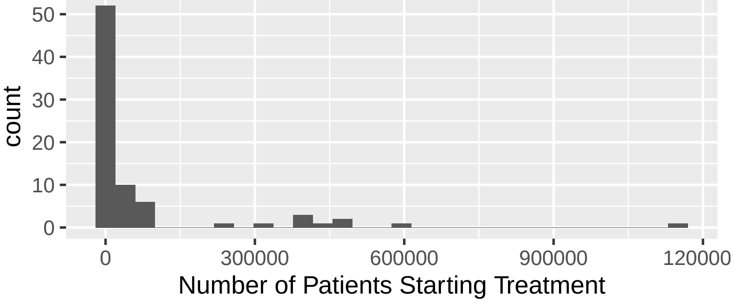
Dion Chang - 20812576 and Anamika Sharma - 20870698

08/02/2021

1) The data set that we worked with was from "Budgets for TB since fiscal year 2018." The raw data set has 645 rows (observations) and 43 columns (variables) with several missing data (N/A) in the cells. The variables we used in this data are the Budget required for TB Preventive treatment in US Dollars (budget_tpt) and Number of patients expected to start TB preventive treatment (tx_tpt). We have also set the year to only look at 2020. To remove any missing values, we have performed an inner join between the variables so both of them in the same row would have a value greater than zero.

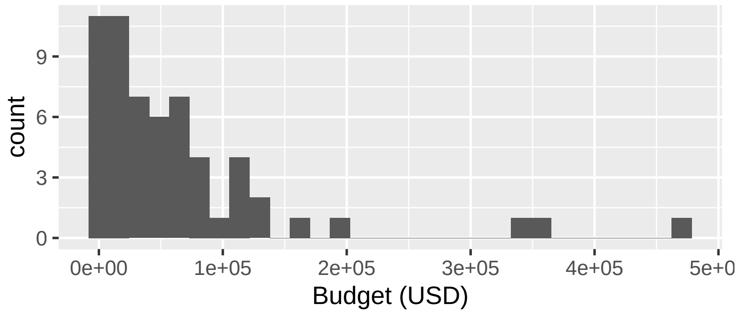
Below are histograms of budget and number of patients starting treatment prior to data cleaning:

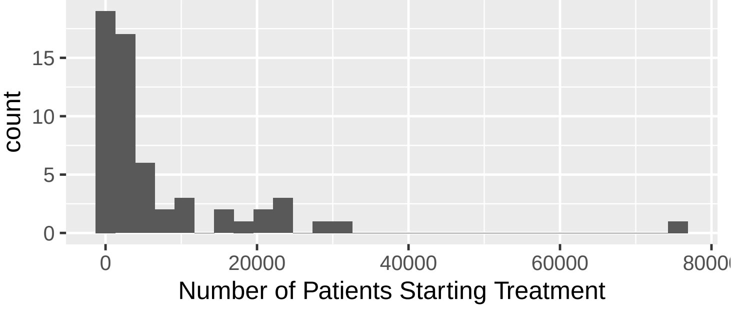




After filtering, removing outliers, and cleaning the data, our data frame has 58 observations/rows and 6 columns. Columns include country, year, budget (before and after log transformation), and number of patients expected to start TB preventive treatment (before and after log transformation). Both variables before log transformation are integer (discrete) data types. After log transformation, the budget and number of patients variables are continuous (numeric). As for summary statistics, the important ones are the following: prior to log transformation, minimum budget was determined to be $2279, maximum budget was $472710. Minimum number of patients was 50, and the maximum number of patients was 75634.

Below are histograms of budget and number of patients starting treatment after filtering, removing outliers, and cleaning the data:



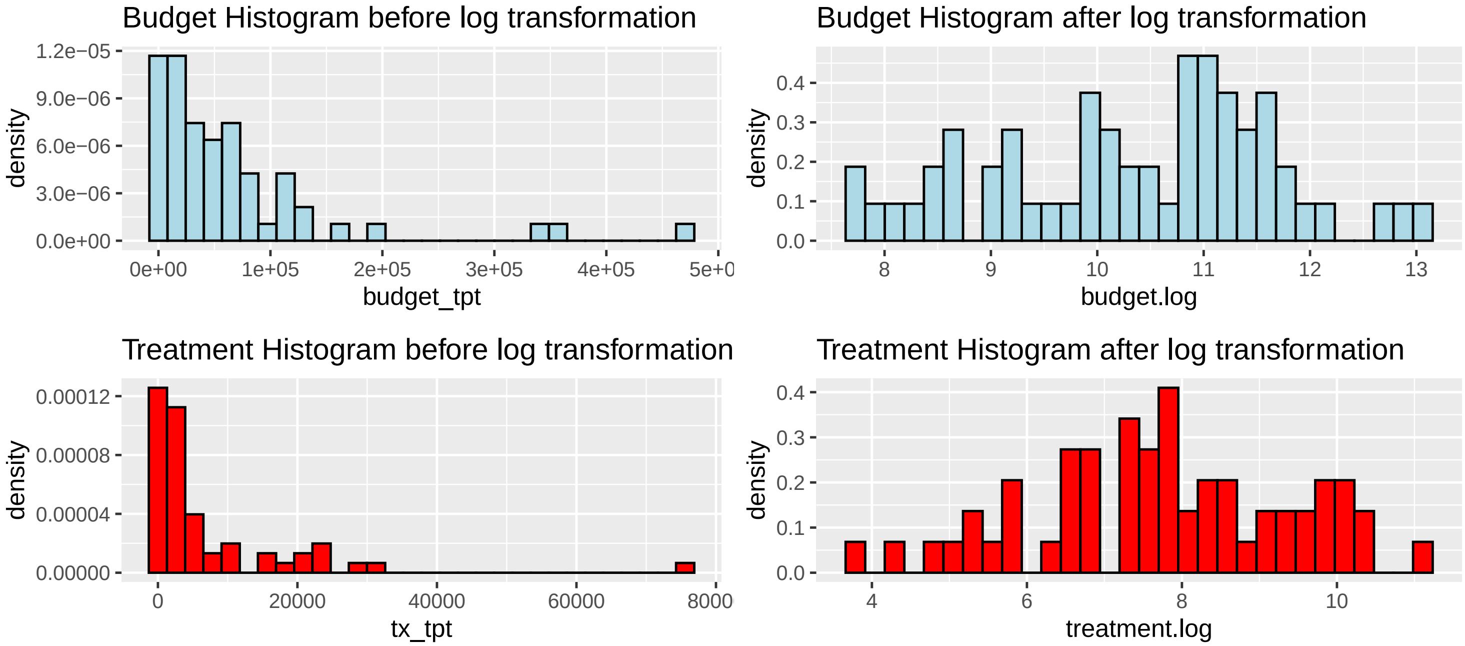


The data overall is positively skewed (skewed to the right) as shown in the histograms. As further supported by the summary statistics, both budget and number of patients (treatments) variables have positive values (2.76 and 3.43 respectively) for skewness. Therefore, we cannot assume normality. More of this is discussed in **2)**.

---

2) The pair of variables that were selected to be analyzed are Budget required for TB preventive treatment in US Dollars (budget_tpt) and Number of patients expected to start TB preventive treatment (tx_tpt).

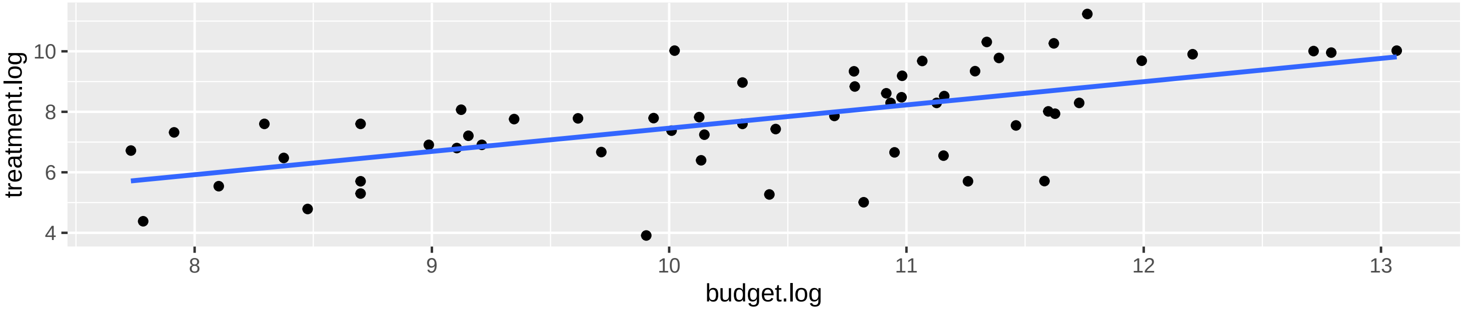
The histograms for Budget and Number of patients expected to start TB treatment below show that we cannot assume normality due to the fact that it was positively skewed (histograms on the left). However, after log transformation, the histograms on the right display normal distribution, which was what we wanted:



We have also plotted the variables using Q-Q plot (shown in the R code file). Squared transformation was also performed, but it was determined that Log transformation had the best performance. Overall, after Log transformation the skewness for budget and number of patients starting treatment have improved to -0.22 and -0.16 (respectively) which is close to zero. This means that they can now be assumed to have normal distribution.

---

3) We believe that there is a positive correlation between budget and number of patients expected to start TB treatment. The correlation plot is displayed below:



The plot shown above overall shows a positive linear (correlation) relationship between budget and number of patients expected to start TB treatment. The assumption of the data was that it is interval data. Therefore, Pearson correlation test was deemed appropriate. The following results from the correlation test are shown below:

correlation coefficient = r = 0.611, $p < 0.001$, 95% CI 0.419 to 0.751

---

4)

From the correlation test, it was determined that Budget required for TB preventive treatment was significantly correlated with Number of patients expected to start TB preventive treatment (r = 0.611, $p < 0.001$, 95% CI 0.419 to 0.751). A correlation of 0.611 represents a large effect explaining ~37% of the variance.

The analysis shows that the correlation between the Budget required for TB preventive treatment and the Number of patients expected to start TB preventive treatment was positive. This means that the greater the budget for tuberculosis preventive treatment, the more patients we can expect to start their tuberculosis preventive treatment. This would be expected.

However, one should also keep in mind that correlation does not imply causation. In this case, a high budget does not suddenly cause an increase number of patients expecting to start tuberculosis treatment.