**Capstone Project 1:**

**Topic: Pump it Up: Data Mining the Water Table.**

**Statistical Data Analysis**

To find the significance of different variable to predict the status of pumps I have conducted the following statistical tests.

**T-test:** A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. The t-test is one of many tests used for the purpose of hypothesis testing in statistics.

The t-test is conducted on two numeric features, amount\_tsh and gps\_height across different status group.

Ho: The average mean of amount\_tsh is same for both functional and nonfunctional group.

H1: The average mean of amount\_tsh is not same for both functional and nonfunctional group.

alpha=.05

if p-value<alpha: reject null hypothesis

if p-value>=alpha: fail to reject null hypothesis

from the test it is identified that both amount\_tsh and gps\_height have p-value less than alpha so, rejected null hypothesis. It is concluded that the mean of amount\_tsh and gps\_height is varying among status\_group. These two features can have important role in the prediction.

**One-Way Anova:** The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. From One-way Anova also, It is concluded that the mean of amount\_tsh and gps\_height is varying among status\_group. These two features can have important role in the prediction.

**Chi-square test:** The Chi Squarestatistic is commonly used for testing relationships between categorical variables.  The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.

A chi-squared test of independence is run to verify the relation between each categorical variable and the Status\_group (dependent) variable. A significance level of 0.05 is chosen. For each test, if the p-value is less than the significance level, the corresponding variable is kept being possibly used in the model.

From the chi\_square test it is identified that,the parameters (water\_quality, extraction\_type, waterpoint\_type, payment, source,source\_class,basin, funder, scheme\_management, management, quantity) have p-value less than alpha so,reject the null hypothesis.That means the above listed features plays a relevant role in the prediction of status\_group.

.