University of Gloucestershire

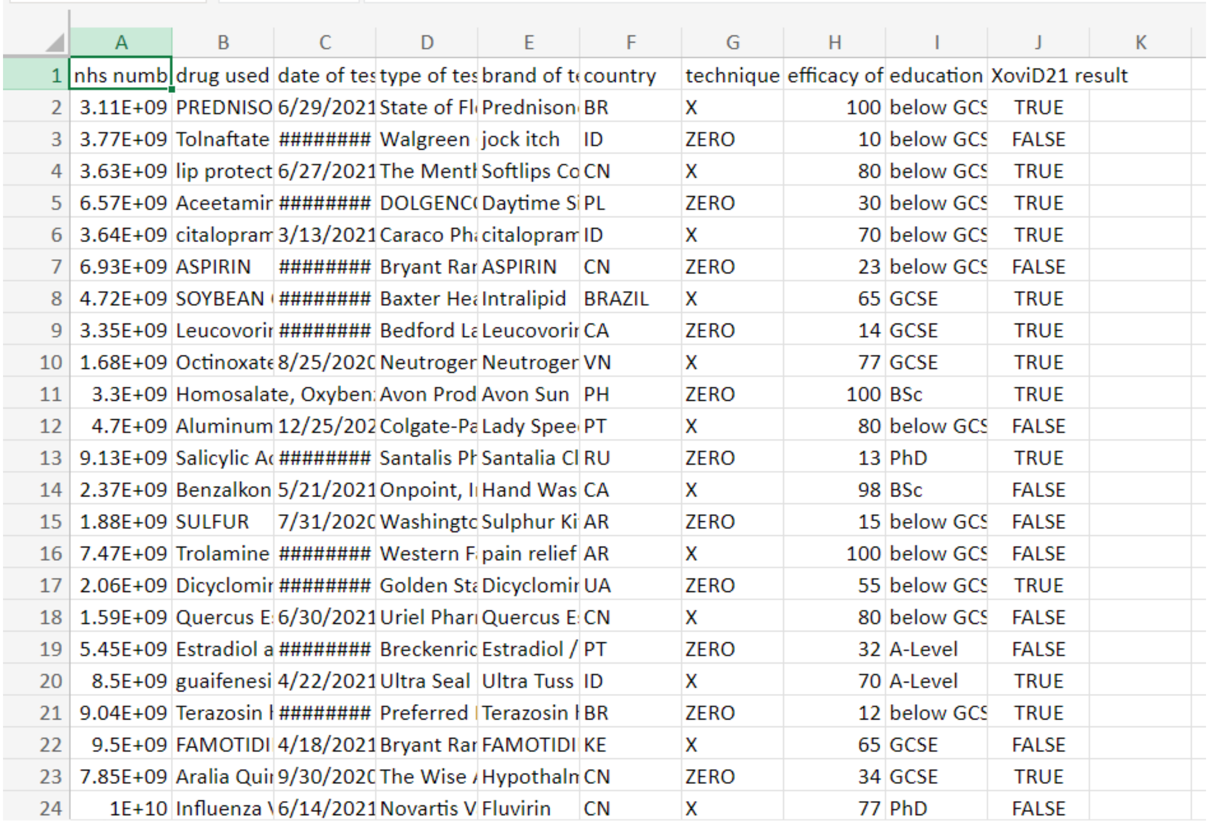
Maths in Data Science (CT4031)

Dataset Analysis, Visualisation & Statistics

Table of Contents

Introduction

This report will cover the accurate complete analysis of the provided dataset related to the different usage of the 2 medical techniques labelled “X” and “ZERO” within drug tests. Additionally, the data will visually be represented and hypothetically tested using Matplotlib – a Python library used for creating “static, animated and interactive” visualisations such as scatter graphs and histograms (Matplotlib, 2012). During this analysis, popular data-analytical techniques which data scientists would implement for a dataset below were evidently revealed within this report in the **Dataset Analysis and Pre-Processing** section. One technique essential for the initiation of the analysis of the dataset below involved the “OSEMN” framework. The cleansing and the preparation for a useful dataset (Kumari, Bhardwaj and Sharma, 2020).



	A	B	C	D	E	F	G	H	I	J	K
1	nhs numb	drug used	date of tes	type of tes	brand of t	country	technique	efficacy of	education	XoviD21 result	
2	3.11E+09	PREDNISO	6/29/2021	State of Fl	Prednison	BR	X	100	below GCS	TRUE	
3	3.77E+09	Tolnaftate	########	Walgreen	jock itch	ID	ZERO	10	below GCS	FALSE	
4	3.63E+09	lip protect	6/27/2021	The Menth	Softlips Co	CN	X	80	below GCS	TRUE	
5	6.57E+09	Aceetamir	########	DOLGENC	Daytime Si	PL	ZERO	30	below GCS	TRUE	
6	3.64E+09	citalopram	3/13/2021	Caraco Ph	citalopram	ID	X	70	below GCS	TRUE	
7	6.93E+09	ASPIRIN	########	Bryant Rar	ASPIRIN	CN	ZERO	23	below GCS	FALSE	
8	4.72E+09	SOYBEAN (	########	Baxter Hea	Intralipid	BRAZIL	X	65	GCSE	TRUE	
9	3.35E+09	Leucovorir	########	Bedford La	Leucovorir	CA	ZERO	14	GCSE	TRUE	
10	1.68E+09	Octinoxate	8/25/2020	Neutroger	Neutroger	VN	X	77	GCSE	TRUE	
11	3.3E+09	Homosalate, Oxyben:		Avon Prod	Avon Sun	PH	ZERO	100	BSc	TRUE	
12	4.7E+09	Aluminum	12/25/202	Colgate-Pa	Lady Spee	PT	X	80	below GCS	FALSE	
13	9.13E+09	Salicylic Ac	########	Santalis Ph	Santalia Cl	RU	ZERO	13	PhD	TRUE	
14	2.37E+09	Benzalkon	5/21/2021	Onpoint, I	Hand Was	CA	X	98	BSc	FALSE	
15	1.88E+09	SULFUR	7/31/2020	Washingtc	Sulphur Ki	AR	ZERO	15	below GCS	FALSE	
16	7.47E+09	Trolamine	########	Western F	pain relief	AR	X	100	below GCS	FALSE	
17	2.06E+09	Dicyclomir	########	Golden Sta	Dicyclomir	UA	ZERO	55	below GCS	TRUE	
18	1.59E+09	Quercus E	6/30/2021	Uriel Phar	Quercus E	CN	X	80	below GCS	FALSE	
19	5.45E+09	Estradiol a	########	Breckenric	Estradiol /	PT	ZERO	32	A-Level	FALSE	
20	8.5E+09	guaifenesi	4/22/2021	Ultra Seal	Ultra Tuss	ID	X	70	A-Level	TRUE	
21	9.04E+09	Terazosin I	########	Preferred	Terazosin I	BR	ZERO	12	below GCS	TRUE	
22	9.5E+09	FAMOTIDI	4/18/2021	Bryant Rar	FAMOTIDI	KE	X	65	GCSE	FALSE	
23	7.85E+09	Aralia Quir	9/30/2020	The Wise /	Hypothaln	CN	ZERO	34	GCSE	TRUE	
24	1E+10	Influenza V	6/14/2021	Novartis V	Fluvirin	CN	X	77	PhD	FALSE	

Figure 1 – Technique” X” and” ZERO” Dataset

The content within Figure 1 shows an unclean dataset which seems to show drugs used on patients, providing their NHS number and education level. It could assume that these patients live in the UK hinted by the presence of one of the columns “nhs number” implying the assumption. Stating the basics, the dataset has 1000 rows of data and 10 columns where multiple drugs are labelled, their brand, the type of drug tests, the 2 techniques mentioned before: “X” and “ZERO”, the efficacy of the drug test and the result Boolean value whether any traces of the COVID-19 virus are still present within the patients. Without performing any changes to the dataset, visible duplicates of incorrect spelling and signs of irrelevant unusual data are common.

Dataset Analysis & Pre-Processing

After dissecting several key trends from the dataset, information was concluded to develop an understanding on the full purpose of the dataset and its usefulness.

The "OSEMN" Framework

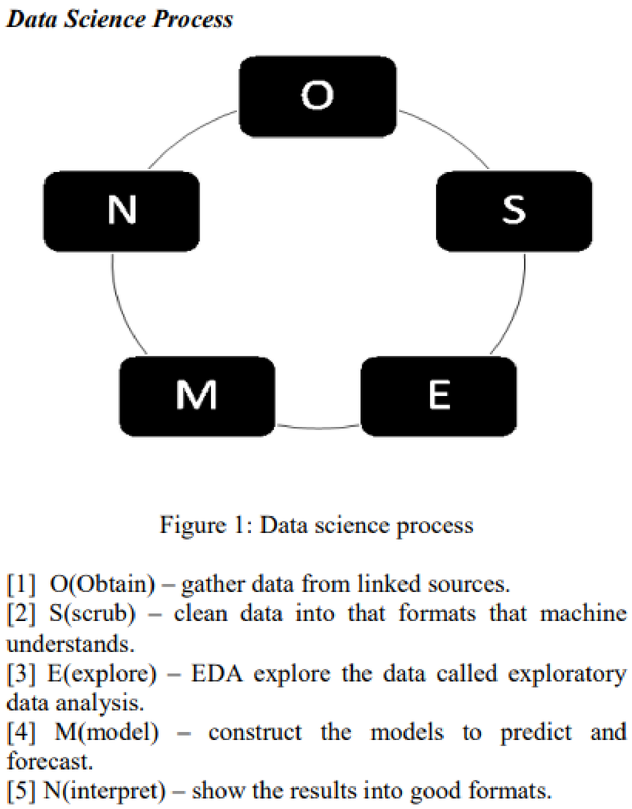


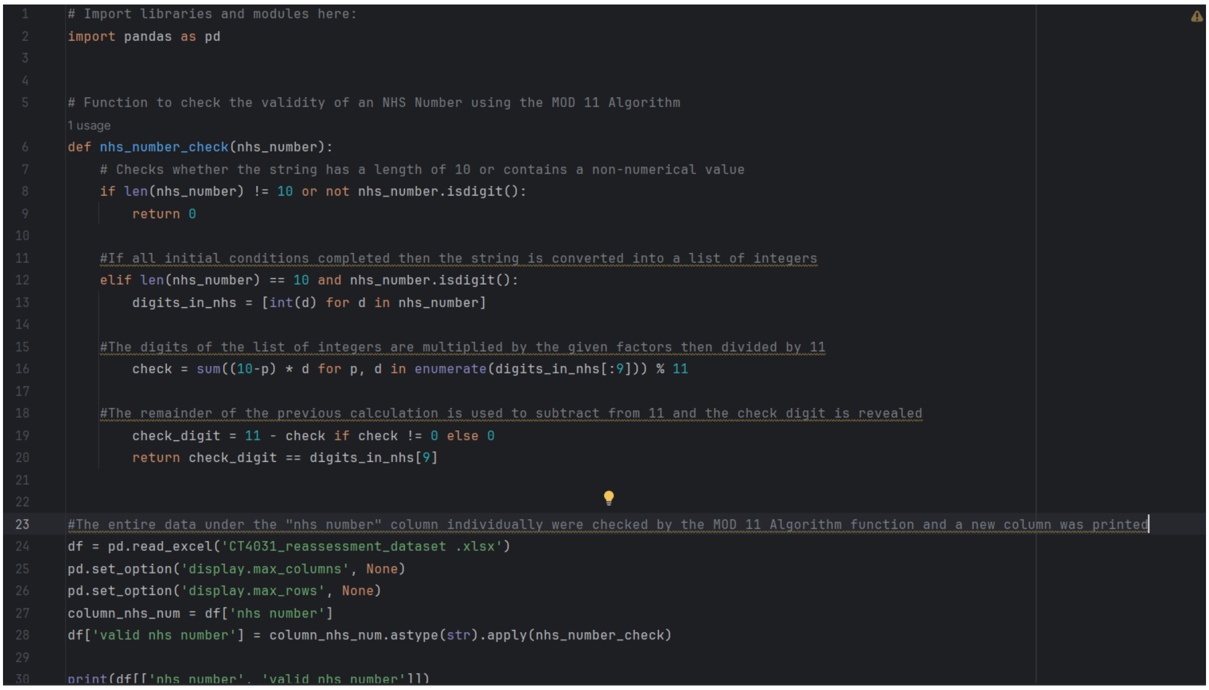
Figure 1: Data science process

[1] O(Obtain) – gather data from linked sources.
[2] S(scrub) – clean data into that formats that machine understands.
[3] E(explore) – EDA explore the data called exploratory data analysis.
[4] M(model) – construct the models to predict and forecast.
[5] N(interpret) – show the results into good formats.

Figure 2 – "OSEMN" Framework (Kumari, Bhardwaj and Sharma, 2020)

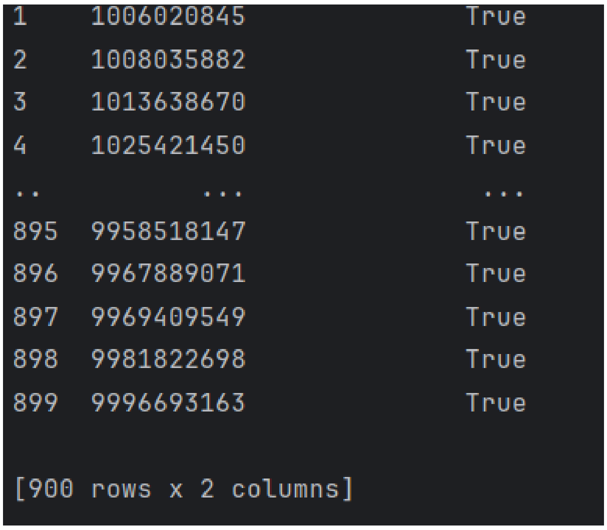
Obtaining the data -> The dataset provided in the excel spreadsheet is the representation of this stage. Data has been collected assumingly from the hospital's database.

Scrubbing the data -> To begin, the data went through a cleaning process to remove unusual activity. The first column, the NHS number was checked. An NHS number is a 10-numerical digit unique identifier used to refer to individuals when using NHS services. The 10th digit at the end is the check digit to confirm validity using the MOD 11 Algorithm (Boyd, Thomas and Macleod, 2018). Using this knowledge, a Python program was produced to confirm the validity of the remaining NHS Numbers. Any set of data under the NHS Number column which had less than 10 digits was automatically removed. 900 sets of data were left.



```python
# Import libraries and modules here:
import pandas as pd


# Function to check the validity of an NHS Number using the MOD 11 Algorithm
1 usage
def nhs_number_check(nhs_number):
    # Checks whether the string has a length of 10 or contains a non-numerical value
    if len(nhs_number) != 10 or not nhs_number.isdigit():
        return 0

    #If all initial conditions completed then the string is converted into a list of integers
    elif len(nhs_number) == 10 and nhs_number.isdigit():
        digits_in_nhs = [int(d) for d in nhs_number]

    #The digits of the list of integers are multiplied by the given factors then divided by 11
        check = sum((10-p) * d for p, d in enumerate(digits_in_nhs[:9])) % 11

    #The remainder of the previous calculation is used to subtract from 11 and the check digit is revealed
        check_digit = 11 - check if check != 0 else 0
        return check_digit == digits_in_nhs[9]


#The entire data under the "nhs number" column individually were checked by the MOD 11 Algorithm function and a new column was printed
df = pd.read_excel('CT4031_reassessment_dataset .xlsx')
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
column_nhs_num = df['nhs number']
df['valid nhs number'] = column_nhs_num.astype(str).apply(nhs_number_check)

print(df[['nhs number', 'valid nhs number']])
```

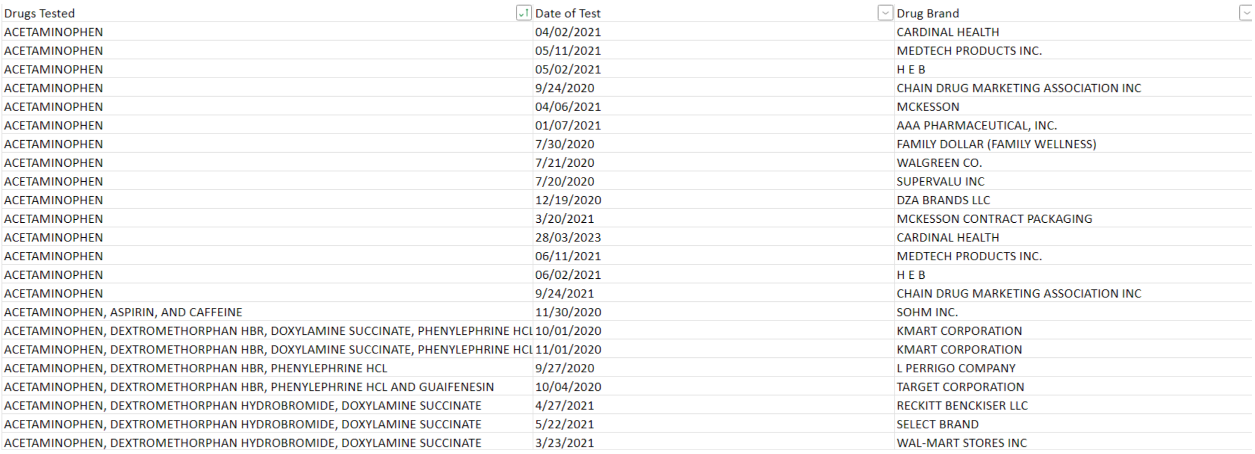
Figure 3– Python Code for NHS Number Validation



```
1    1006020845              True
2    1008035882              True
3    1013638670              True
4    1025421450              True
..          ...               ...
895  9958518147              True
896  9967889071              True
897  9969409549              True
898  9981822698              True
899  9996693163              True

[900 rows x 2 columns]
```

Figure 4 – Python Code Results for Dataset's NHS Numbers

Each column was properly named for organisation and to better understand what the data is showing. For example, the column "country" was renamed "Country of Brand" as the former column name was vague. Looking through much of the data within our columns, it was simpler to standardize all the data. Within the date column, the Month-Date-Year format was in place (given leading zeros), the drugs, it's brand and the drug type were all capitalized as the standard format. All the numerical values of the efficacy of the drug test technique were rounded to the nearest 1th. All duplicates or incorrectly spelt data were either completely removed or assumingly modified. This was achieved by using several data sorting features embedded within Microsoft Excel such as the filter feature. 850 pieces of data were left.



Drugs Tested	Date of Test	Drug Brand
ACETAMINOPHEN	04/02/2021	CARDINAL HEALTH
ACETAMINOPHEN	05/11/2021	MEDTECH PRODUCTS INC.
ACETAMINOPHEN	05/02/2021	H E B
ACETAMINOPHEN	9/24/2020	CHAIN DRUG MARKETING ASSOCIATION INC
ACETAMINOPHEN	04/06/2021	MCKESSON
ACETAMINOPHEN	01/07/2021	AAA PHARMACEUTICAL, INC.
ACETAMINOPHEN	7/30/2020	FAMILY DOLLAR (FAMILY WELLNESS)
ACETAMINOPHEN	7/21/2020	WALGREEN CO.
ACETAMINOPHEN	7/20/2020	SUPERVALU INC
ACETAMINOPHEN	12/19/2020	DZA BRANDS LLC
ACETAMINOPHEN	3/20/2021	MCKESSON CONTRACT PACKAGING
ACETAMINOPHEN	28/03/2023	CARDINAL HEALTH
ACETAMINOPHEN	06/11/2021	MEDTECH PRODUCTS INC.
ACETAMINOPHEN	06/02/2021	H E B
ACETAMINOPHEN	9/24/2021	CHAIN DRUG MARKETING ASSOCIATION INC
ACETAMINOPHEN, ASPIRIN, AND CAFFEINE	11/30/2020	SOHM INC.
ACETAMINOPHEN, DEXTROMETHORPHAN HBR, DOXYLAMINE SUCCINATE, PHENYLEPHRINE HCL	10/01/2020	KMART CORPORATION
ACETAMINOPHEN, DEXTROMETHORPHAN HBR, DOXYLAMINE SUCCINATE, PHENYLEPHRINE HCL	11/01/2020	KMART CORPORATION
ACETAMINOPHEN, DEXTROMETHORPHAN HBR, PHENYLEPHRINE HCL	9/27/2020	L PERRIGO COMPANY
ACETAMINOPHEN, DEXTROMETHORPHAN HBR, PHENYLEPHRINE HCL AND GUAIFENESIN	10/04/2020	TARGET CORPORATION
ACETAMINOPHEN, DEXTROMETHORPHAN HYDROBROMIDE, DOXYLAMINE SUCCINATE	4/27/2021	RECKITT BENCKISER LLC
ACETAMINOPHEN, DEXTROMETHORPHAN HYDROBROMIDE, DOXYLAMINE SUCCINATE	5/22/2021	SELECT BRAND
ACETAMINOPHEN, DEXTROMETHORPHAN HYDROBROMIDE, DOXYLAMINE SUCCINATE	3/23/2021	WAL-MART STORES INC

Figure 5 – Partially Standardized Dataset

Exploring the data -> The next phase was to use statistical quantities and techniques to explore the data and gather valuable results to interpret these results and provide a clear useful conclusion. By using the data under the "Country of Brand", "Drug Test Technique", "Efficacy of Technique" and "XoviD21 Result", several trends relating to the drugs were produced. The following trends were described:

- 449 sets of data showed that patients tested positive for COVID-19 where 207 unique drugs were shown, with the top six drugs being "Alcohol" (or any other alcohols), "Acetaminophen", "Oxygen", "Oxtinoxate", "Triclosan" and "Diltiazem Hydrochloride" where the Chinese drug brand "RemedyRepack Inc" takes the top.

- Technique "ZERO" shows to be the recurring technique present for the positive COVID-19 results where the average efficacy was 37%

- On the other hand, 401 sets of data showed patients tested negative for COVID-19 where 198 unique drugs were shown, with the top six being the "Chloride" drugs, "Acetaminophen", "Oxtinoxate", "Alcohol", "Hydrocodone Bitartrate" and "Triclosan" where the Japanese drug brand "Nelco Laboratories Inc" leads.

- Technique "ZERO" seems to be the recurring technique present for the negative COVID-19 results where the average efficacy is 40%

- 400 sets of data use the technique "X" where the average efficacy is 73% while 450 sets of data use the technique "ZERO" where the average efficacy is 39% (when patient's COVID-19 results are positive and negative)

Data Visualisation of the Dataset

Visuals of some of the trends mentioned beforehand within **Dataset Analysis and Pre-Processing,** calculations of statistical quantities were required such as the mean, median, mode, range, minimum, maximum and the standard deviation. Notably the efficacy of the two different drug test techniques ZERO and X to compare these techniques and build hypotheses around these statistics. The visualisation of the data is shown below:

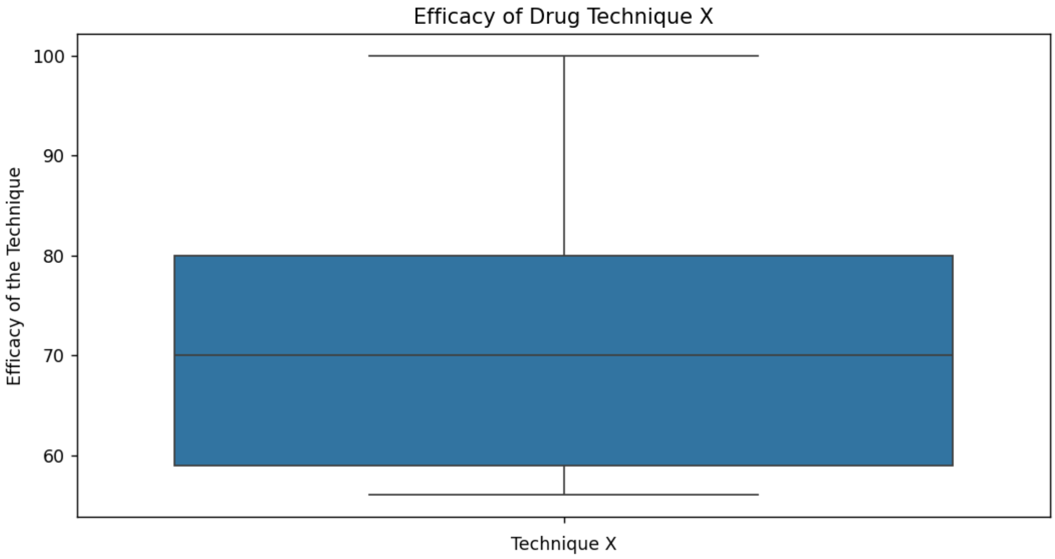


Figure 6 – Box Plot of Efficacy of Technique "X"

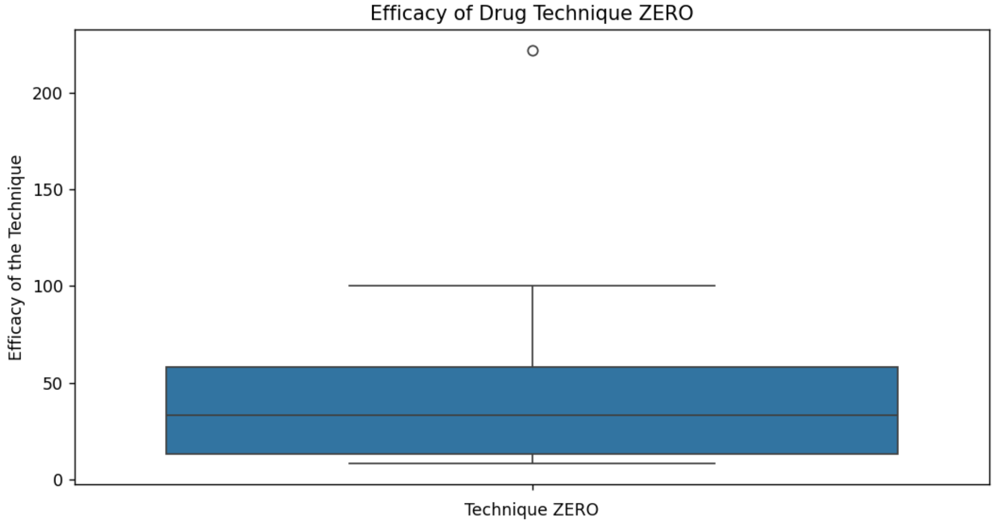


Figure 7 – Box Plot of Efficacy of Technique "ZERO"

Figure 6 and 7 both show box plots of the two different drug techniques "X" and "ZERO". From these two box plots, both statistical quantities from the box plots can be compared to deduce useful conclusions for medical research. Both techniques share a maximum efficacy of 100 therefore at least several attempts of using both techniques have been highly successful however this does not prove that both techniques are equal in terms of average efficacy. This is shown by the difference in minimum value, the median and the quartiles of the different box plots. The minimum value, median and quartiles of technique "ZERO" are lower than the minimum value of technique "X". This shows that "ZERO" is less effective compared to "X". It could be inferred that "X" is more likely to be effective at providing a false result testing for COVID-19 however these data visuals do not prove any correlation between efficacy of the techniques and the COVID-19 results.

Hypothesis Testing of the Dataset

Recently, a hypothesis was generated that there is a "statistically significant difference between Technique "X" and "ZERO". Beforehand, we found a trend that technique "X" is more effective than "ZERO" in terms of efficacy however other statistical data needs to be considered such as the COVID-19 test results. Below is the complete hypothesis test:

Null Hypothesis: "There is no statistically significant difference between Technique X and ZERO"

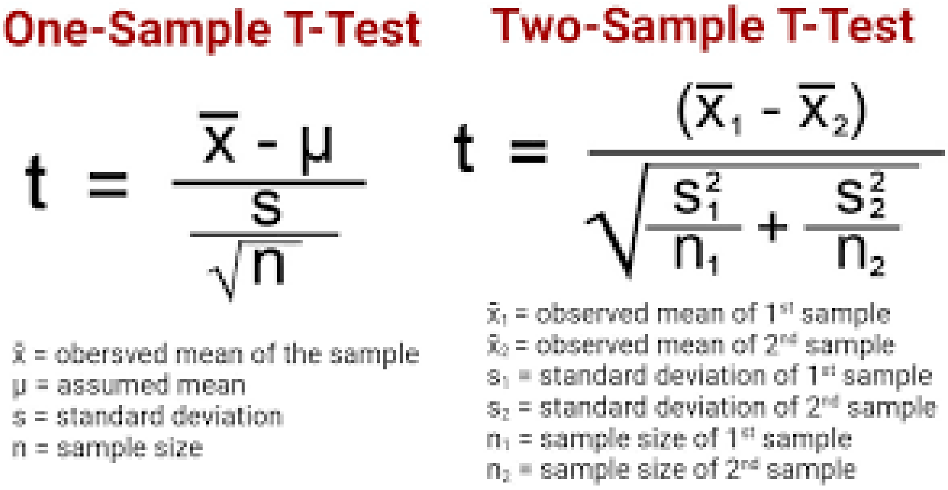
Alternate Hypothesis: "There is a statistically significant difference between Technique X and ZERO"

The level of significance used will be 0.05. Generally, it is labelled that a p-value is less than 0.05 is it deemed as statistically significant. A t-test will be carried out using the means and standard deviations of the efficacies of the two techniques analysed from the box plots.

The mean efficacy of X: 72.9 & the mean efficacy of ZERO: 38.4

The standard deviation of X: 14.7 & the standard deviation of ZERO: 26.9

Sample size of X: 400 & sample size of ZERO: 449



One-Sample T-Test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{x}$ = obersved mean of the sample
μ = assumed mean
s = standard deviation
n = sample size

Two-Sample T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = observed mean of 1st sample
$\bar{x}_2$ = observed mean of 2nd sample
s_1 = standard deviation of 1st sample
s_2 = standard deviation of 2nd sample
n_1 = sample size of 1st sample
n_2 = sample size of 2nd sample

The T-test value is 22.65 where the p-value is under the given test significance of 0.05 therefore the null hypothesis is rejected and there is a statistically significant difference between the two techniques.

Conclusion

The dataset has been analysed and it has shown us a trend of Technique X as the significantly more effective technique suggested to use for medical purposes.

Reference List

Matplotlib (2012). Matplotlib: Python plotting — Matplotlib 3.1.1 documentation. [online] Matplotlib.org. Available at: https://matplotlib.org/

Kumari, K., Bhardwaj, M. and Sharma, S. (2020). OSEMN Approach for Real Time Data Analysis. International Journal of Engineering and Management Research, 10(02), pp.107–110. doi:https://doi.org/10.31033/ijemr.10.2.11.

Boyd A, Thomas R, Macleod J. NHS Numbers and their management systems. London, UK: CLOSER; 2018