Scholastica Olanrewaju

MPH, PMP

HEALTH DATA ANALYST AND PROJECT MANAGER

Hello!
I am Scholastica

I have worked in health data analysis, health consultancy and program implementation for the past 8 years. I am passionate about drawing out insights from data and translating same in decision-making for positive impact while ensuring efficient use of resources.

I have a graduate degree in Master of Public Health (University of Aberdeen), nanodegree in data analysis (Udacity) and I am a certified Project Management Professional (PMP)

Career Timeline

2011

Graduated with **BSc (First Class)** from University of Ibadan

2012

Served* in Institute of Human Virology Nigeria (IHVN), was retained as a full staff, and promoted in its Strategic Information Department**

2018

Graduated with a **Master of Public Health (MPH)** from University of Aberdeen, UK as a Commonwealth Scholar

2019

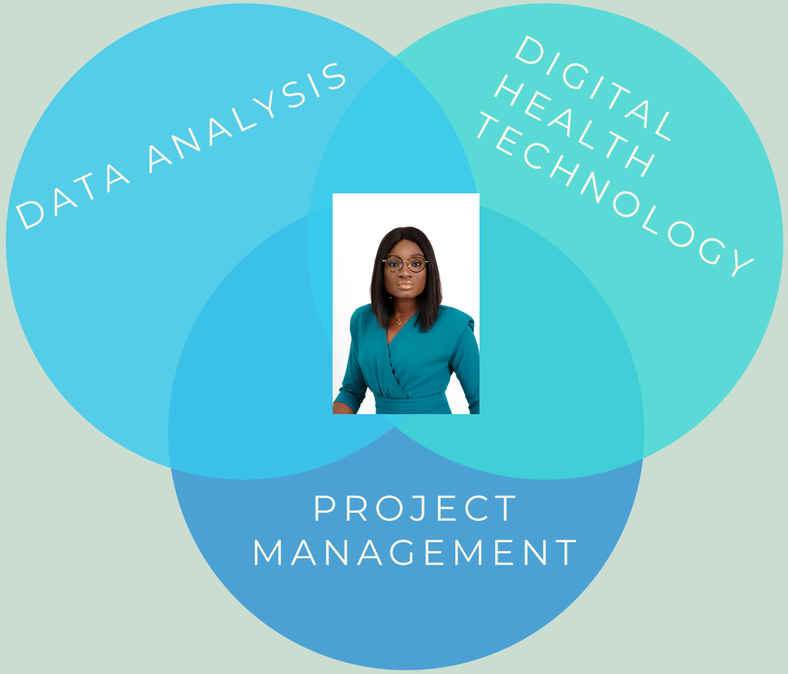
Joined the Strategic Information Department** of Center for Integrated Health Programs (CIHP) and was promoted in 6 months

2020

Became a certified Project Management Professional **(PMP)**

Re-joined IHVN's Strategic Information Department** as a Senior Program Officer

Areas of Interest



Relevant Courses

Data Analyst Nanodegree (2020)
Udacity

The Data Scientist's Toolbox (2021)
Coursera

R Programming (2021)
Coursera

Advanced Methods in Global Health (2021)
Barcelona Institute of Global Health

Technical Skills

*Performed the mandatory 1-year National Youth Service

**Strategic Information Department is responsible for managing the program's data- data collection, review, analysis, communication and use. As well as, the development and maintenance of electronic health records and other component of the health information system.

Overview of Select Data Projects

Explore Weather Trends

An analysis of global and local weather trends between 1856 and 2013

Investigate a Dataset

An analysis of the no-show appointments dataset to identify factors that predicts if a patient will show-up for scheduled appointment

Analyse A/B Test Result

Conducted experiment(s) and analysed data to help an e-commerce company make a decison regarding its webpage

Wrangle and Analyse Data

Wrangle WeRateDogs Twitter data by querying Twitter's API to create interesting and trustworthy analyses and visualizations.

Communicate Data Findings

An analysis to assess factors that affect borrower's APR for the loans.

Develop an R package

Develop an R package to simplify routine analysis of the Nigeria National Data Repository (NDR)"

Explore Weather Trends

INTRODUCTION

In this project, I analysed data on average temperature in Abuja and globally between 1856 and 2013. The **goal** of this analysis is "to create a visualization and prepare a write up describing the similarities and differences between global temperature trends and temperature trends in the closest big city to where you live."

MY APPROACH

This analysis was done systematically using 5-step approach as shown in the figure below

Question(s)

Ensured a good understanding of the questions to guide the analysis

Wrangle

Gathered, assessed and cleaned data using SQL queries

Explore

Explored the data to get clarity of the dataset and maximise the potential of the analysis

Draw Conclusions

Used descriptive statistics (moving average) and visualisations to make observation on the dataset. Intepreted the result

Communicate

Communicated insights, visualisations and conclusions in a written report

ANALYSIS TOOLS

Explore Weather Trends

Input HISTORY MENU

SCHEMA

city_data

city_list

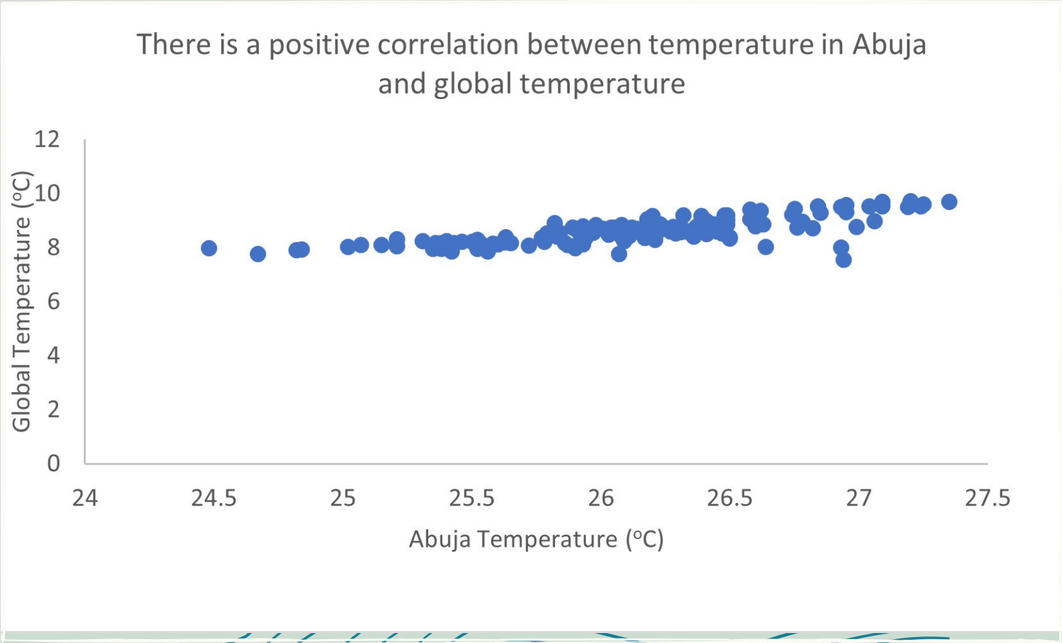
global_data

```
1  SELECT cd.*, gd.year AS global_year, gd.avg_temp AS
   global_temp
2  FROM city_data AS cd
3  JOIN global_data AS gd
4  ON cd.year = gd.year
5  WHERE cd.city = 'Abuja'
```

Success! EVALUATE

Output 158 results Download CSV

year	city	country	avg_temp	global_year	global_temp
1856	Abuja	Nigeria	26.93	1856	8.00
1857	Abuja	Nigeria	24.67	1857	7.76



CONCLUSIONS

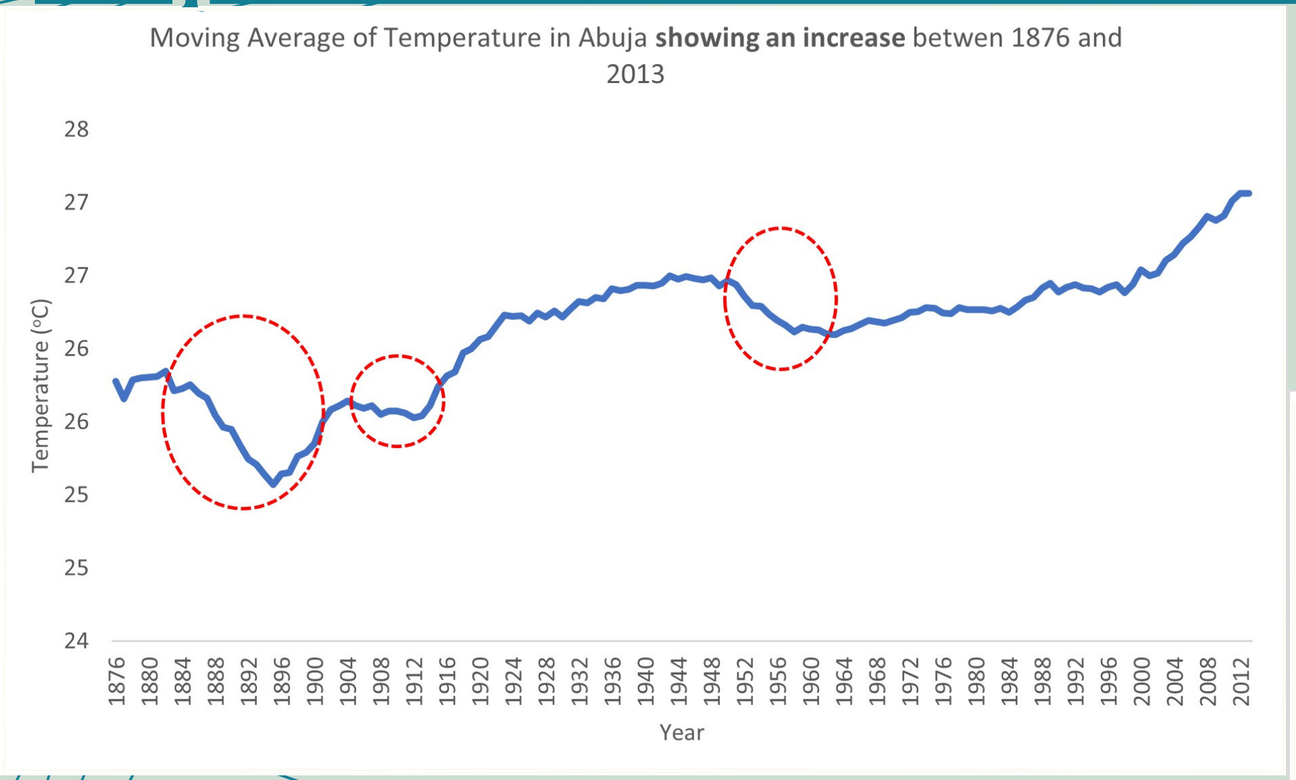
By applying descriptive analysis and visualisation of the data, the following conclusions was drawn in response to the questions.

1. While temperature in Abuja is **normally distributed**, the global temperature is **positively skewed.**

2. The scatter-plot and correlation coefficient of **0.76** indicates a **positive correlation** implying that as global temperature is increasing, the temperature in Abuja is also increasing. Note that this does not imply causation.

3. On average, Abuja is consistently **hotter than the global temperature**. Abuja's temperature was an average of 26.1 (range: 24- 26) degree Celcius while the global temperature average was 8.6 (range: 7-9) degree Celcius within the period of review.

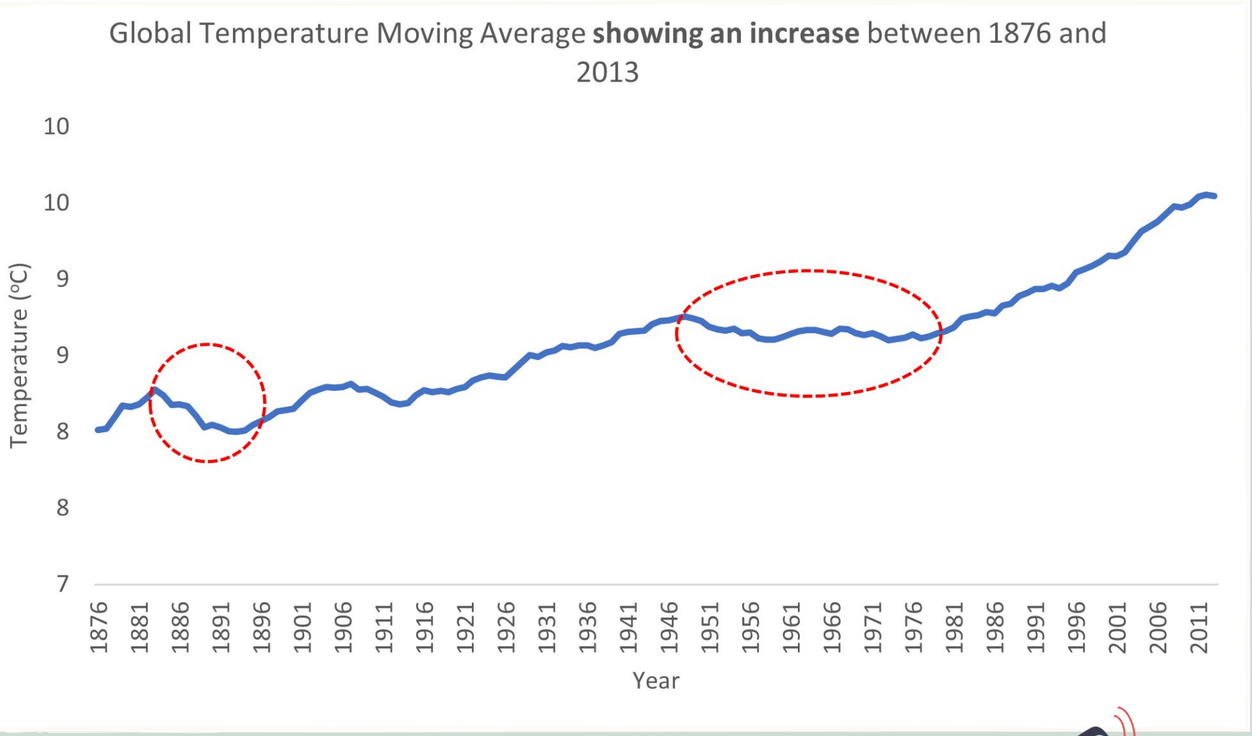
Explore Weather Trends



CONCLUSIONS

Overall, within the period of review (1876 to 2013), the temperature in both Abuja and the Globe showed an **upward trend.** That is, Abuja and the globe is getting hotter. However, there were years within the period when the temperature **showed a decline.** For example- between 1886 -1895 and between 1955-1958.

4



This analysis showed that while the temperature in Abuja is higher than the global temperature, the temperature in Abuja and the globe are positively correlated and increased between the period of review.

It is recommended that research should be done on the period 1886-1895 and 1955-1958 when the temperature declared as there might be lessons from history in it to guide relevant stakeholders to stop global warming.

Investigate a Dataset

INTRODUCTION

In this project, I analysed the *no-show appointments* dataset to identify factors that predicts if a patient will show-up for scheduled appointment. The analysis aimed to answer the following questions:

1. What factors are important to predict if a patient will show up for scheduled appointment?
2. Is duration (appointment day - scheduled day) associated with no-show?

ANALYSIS TOOLS

Pandas, Numpy and **Matplotlib** /ibraries were used in **Python.**

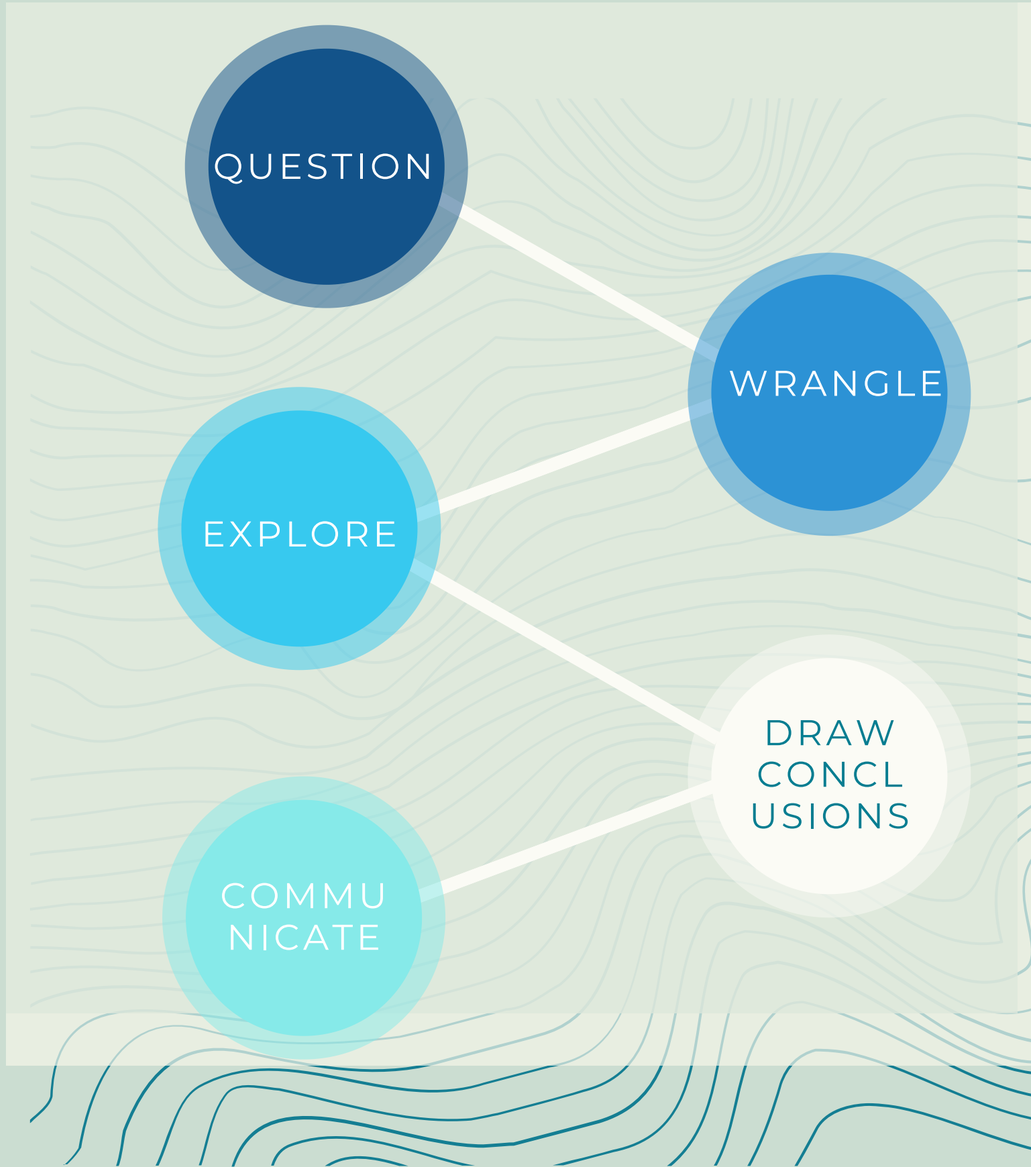
CONCLUSIONS

In response to the analysis questions posed, the findings indicate that:

1. ***age, gender*** **and** ***scholarship*** are important to predict if a patient will show for appointment.
2. Yes, **duration is associated with '*no-show*'**. More patients showed for appointments with a duration that is shorter than 20 days.

Based on the results from this analysis, it is recommended that ***shorter duration be given to patients*** **seeking appointments**. Also, **SMS or email reminder should be considered and targeted to patients** who are more likely to miss appointments. These includes those 10+ years old, females and people not on scholarship.

MY APPROACH



Investigate a Dataset

jupyter Investigate a Dataset Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Investigate a Dataset- Identifying Factors that Predict Missed Appointment Among Patients in Brazil

Table of Contents

Introduction

In this project, I analysed the *no-show appointments* dataset to identify factors that predicts if a patient will show-up for schedule

Dataset

The *No-show appointments* dataset collects information from 100,000 medical appointments in Brazil. A number of characteristic included in each row. The variable included in the dataset are *patient ID, appointment ID, gender, scheduled day, appointment da*

jupyter Investigate a Dataset Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

a. Is age associated with no-show?

Age is numerical and skewed, hence the best descriptive statistics will be median. Younger people missed more appointments when compared against the age of those who show-up for appointment. Overall sample size per age might have play a role in this. Compared with other ages, those less than 10 show-up for most appointments.

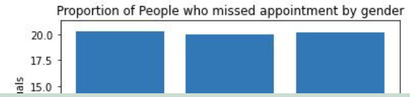
for crosstab

In [24]:
```
# df.groupby('diabetes').showed.median() #this didnt work

# this didn't work. All efforts to troubleshot it - including changing datatype, failed.
```

In [25]:
```
# frequency table for gender by no-show

summary_gender = pd.crosstab(index=df['gender'], columns=df['no_show'], margins =True)
summary_gender['proportion'] = summary_gender['Yes']/summary_gender['All']*100
print(summary_gender)
```

```
no_show     No    Yes     All  proportion
gender
F        57245  14591   71836   20.311543
M        30962   7723   38685   19.963810
All      88207  22314  110521   20.189828
```

In [26]:
```
# Create a bar chart
locations = [1, 2, 3]
heights = summary_gender['proportion']
labels = ['Female', 'Male', 'All']
plt.bar(locations, heights, tick_label=labels)
plt.title('Proportion of People who missed appointment by gender')
plt.xlabel('Gender')
plt.ylabel('Number of Individuals');
```



RELEVANT LINKS

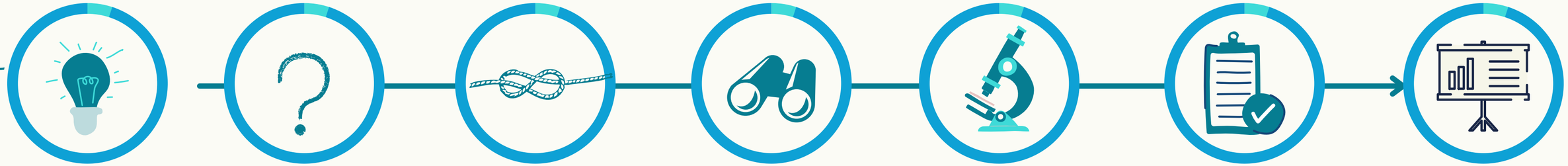
1. View full report and code in Jupyter Notebook
2. Download data set

Analyze A/B Test Result

INTRODUCTION

In this project, I analysed the data to understand the results of an **A/B test** run by an e-commerce website. The **conversion rate** was the metric used for this analysis. The **goal** is: *'to conduct experiment(s) and analyse data to help the company understand if they should implement the new page, keep the old page, or perhaps run the experiment longer to make their decision.'*

MY ANALYSIS APPROACH

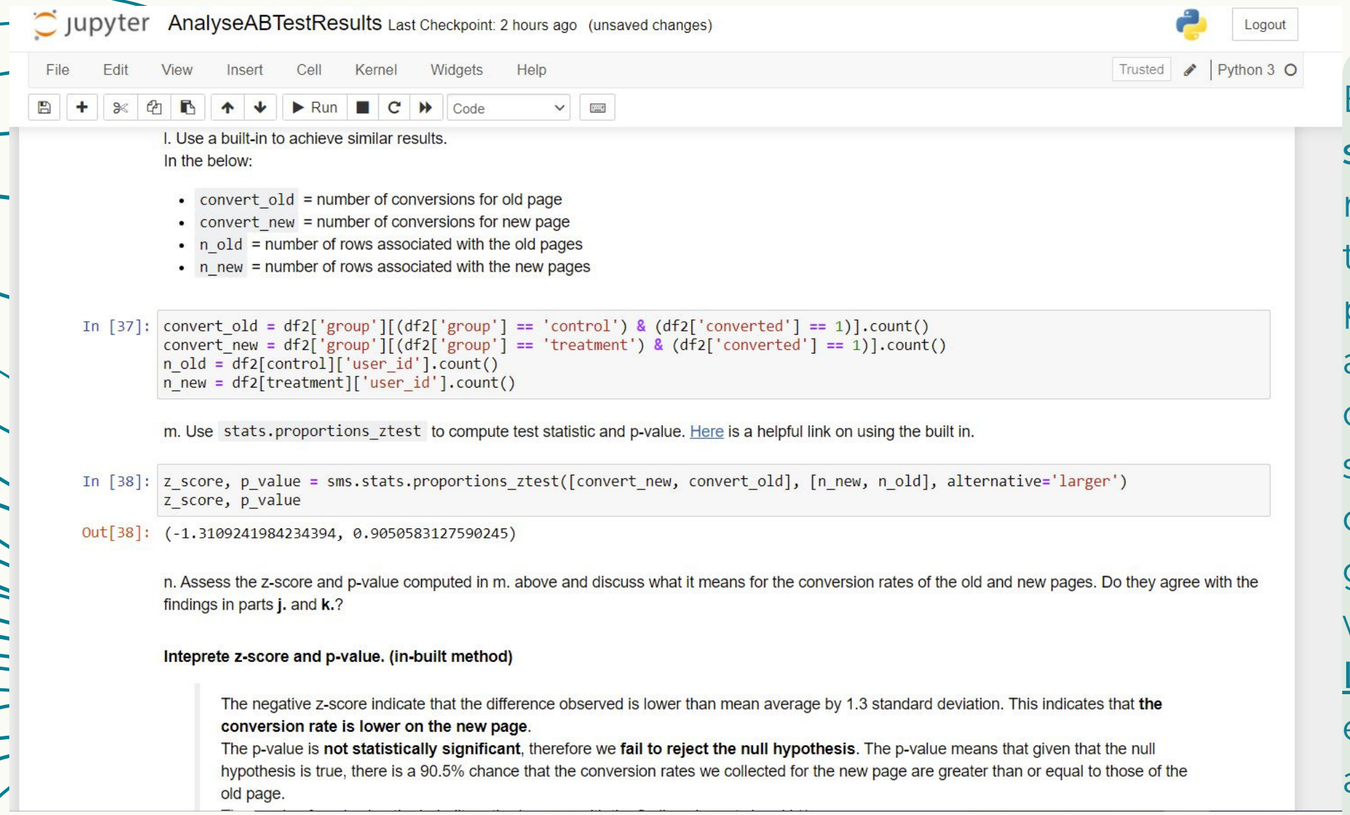


Understand dataset and problem → Hypothesise → Wrangle → Explore → Test Hypothesis → Draw Conclusions → Communicate

ANALYSIS TOOLS

Pandas, Numpy, Matplotlib and **statmodels.api** libraries were used in **Python.**

Analyze A/B Test Result



jupyter AnalyseABTestResults Last Checkpoint: 2 hours ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

l. Use a built-in to achieve similar results.
In the below:

- `convert_old` = number of conversions for old page
- `convert_new` = number of conversions for new page
- `n_old` = number of rows associated with the old pages
- `n_new` = number of rows associated with the new pages

In [37]:
```
convert_old = df2['group'][(df2['group'] == 'control') & (df2['converted'] == 1)].count()
convert_new = df2['group'][(df2['group'] == 'treatment') & (df2['converted'] == 1)].count()
n_old = df2[control]['user_id'].count()
n_new = df2[treatment]['user_id'].count()
```

m. Use `stats.proportions_ztest` to compute test statistic and p-value. Here is a helpful link on using the built in.

In [38]:
```
z_score, p_value = sms.stats.proportions_ztest([convert_new, convert_old], [n_new, n_old], alternative='larger')
z_score, p_value
```

Out[38]: (-1.3109241984234394, 0.9050583127590245)

n. Assess the z-score and p-value computed in m. above and discuss what it means for the conversion rates of the old and new pages. Do they agree with the findings in parts **j.** and **k.**?

Intepret z-score and p-value. (in-built method)

> The negative z-score indicate that the difference observed is lower than mean average by 1.3 standard deviation. This indicates that **the conversion rate is lower on the new page**.
> The p-value is **not statistically significant**, therefore we **fail to reject the null hypothesis**. The p-value means that given that the null hypothesis is true, there is a 90.5% chance that the conversion rates we collected for the new page are greater than or equal to those of the old page.

RELEVANT LINKS

1. View full report and code in Jupyter Notebook

CONCLUSIONS

Based on the results of this analysis, the p-value is **not statistically significant** both in the A/B testing and regression. Therefore, there is **no evidence** to suggest that the new page will lead to more conversion than the old-page. In fact, if an individual landed on the new page, they are **1.015 less likely** to be converted than than if they landed on the old page, holding all variables constant. This is also supported by result from the probability rate with a difference of **-0.001** between the treatment and control group. A deep-dive into available data indicated that there was no evidence of that country affects conversion rate.

Limitation: I could not run an analysis to assess if the experiment was run long enough to cancel novelty effect and change aversion as date was not included in the dataset. Therefore, I recommend further analysis to determine the duration of the experiments and additional data on type of visitors (new or returning).

Overall, I recommend that the e-commerce company ***should run the experiment longer before making a final decision.***

Wrangle and Analyse Data

INTRODUCTION

In this project, I wrangled, analysed and visualised data from the Twitter user- @dog_rates, also known as **WeRateDogs.** WeRateDogs is a Twitter account that rates people's dogs based on pictures or videos with a humorous comment about the dog. The dogs are rated over 10 although the numerator can be above 10. WeRateDogs has over 4 million followers and has received international media coverage.

The **goal** of this project is to '**wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.**'

Python, Excel and Jupyter Notebook was used for this analysis.

DATA SET

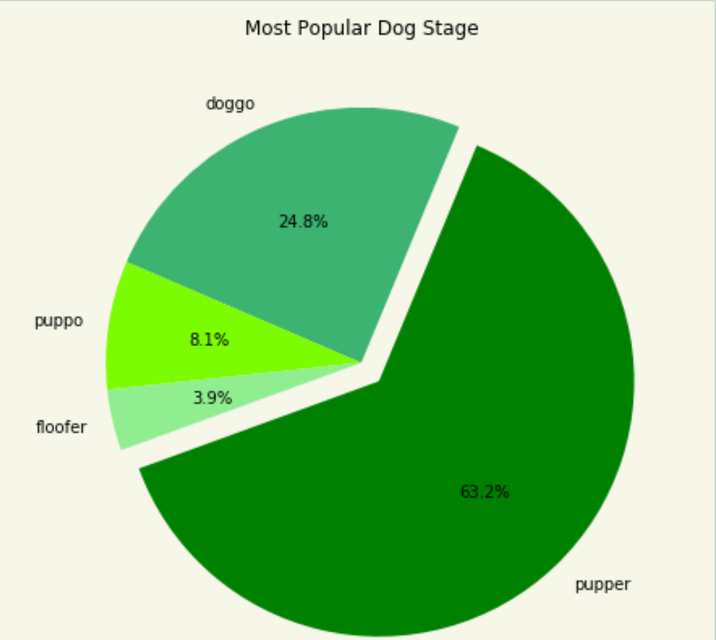
Three dataset was used for this analysis

1. *twitter-archive-enhanced.csv* - this had tweet id and rating data for each dog from the WeRateDogs Twitter Archive*
2. *image_predictions.tsv*- this contains data about the breed of dogs as classified by a running the twitter archive data through a neural network.
3. *additional data from Twitter* - this was gathered programmatically by **querying Twitter's API**. It includes data about retweets, likes etc, of the tweets from WeRateDogs

ANALYSIS TOOLS

Pandas, Numpy, Matplotlib, Tweepy, Request, Seaborn and **Datetime** libraries were used in **Python.**

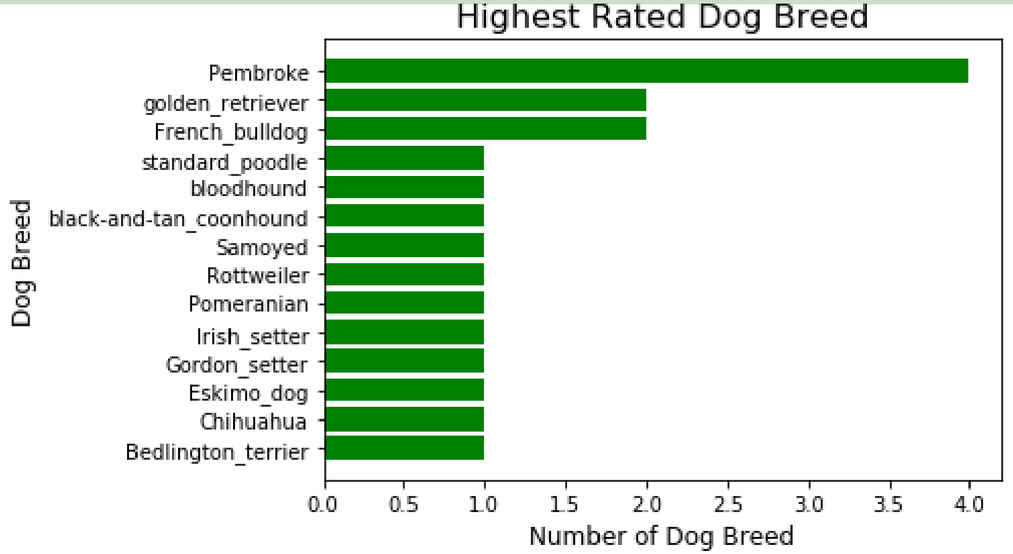
FINDINGS



Pupper' is the most popular dog stage

Wrangle and Analyse Data

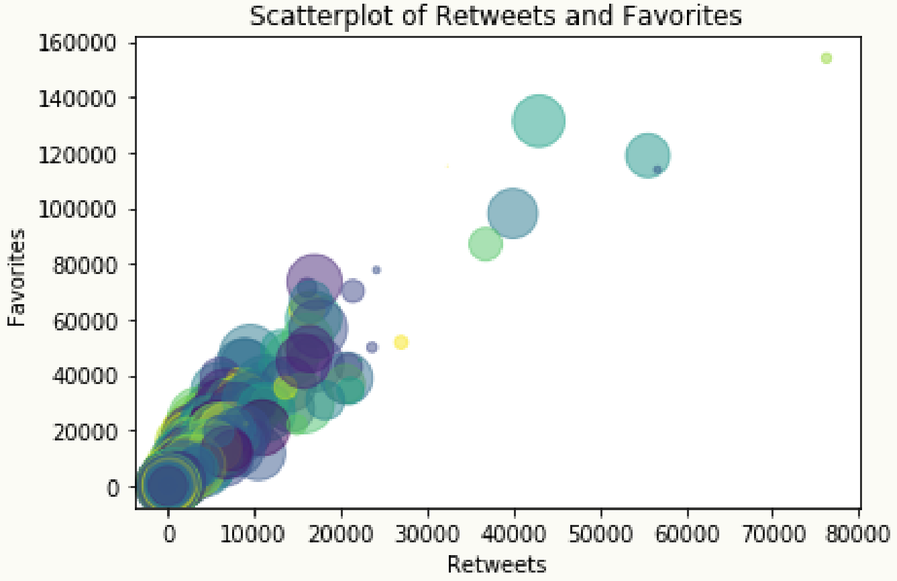
FINDINGS



*The **Bedlington Terrier** is the lowest rated dog breed while the highest rated dog breed is **Pembroke***

Overall, the unnamed dog in the tweet - *'Here's a doggo realizing you can stand in a pool. 13/10 enlightened af'* is clearly the MVP with the highest retweets ane likes. This could be tied to the fact that the tweet includes a video showing the doggo standing comfortable in water- a feat only the brave undertake!

- Retweets and likes are positively correlated.
- 4/5 of the most retweeted were videos of dogs participating in 'fun' activities such as standing on water, blowing bubbles in a bowl of water etc



CONCLUSION

If you are aiming for high number of retweets/likes for a dog-related tweet, a video of the dog in a water-related activity is a winner.

RELEVANT LINKS

1. Read full report here
2. View jupyter notebook which contains code and the analysis process.
3. View data wrangling report

Communicate Data Findings

Effect Loan Characteristics on Borrower's APR

Scholastica Olanrewaju

Click to view presentation

INTRODUCTION

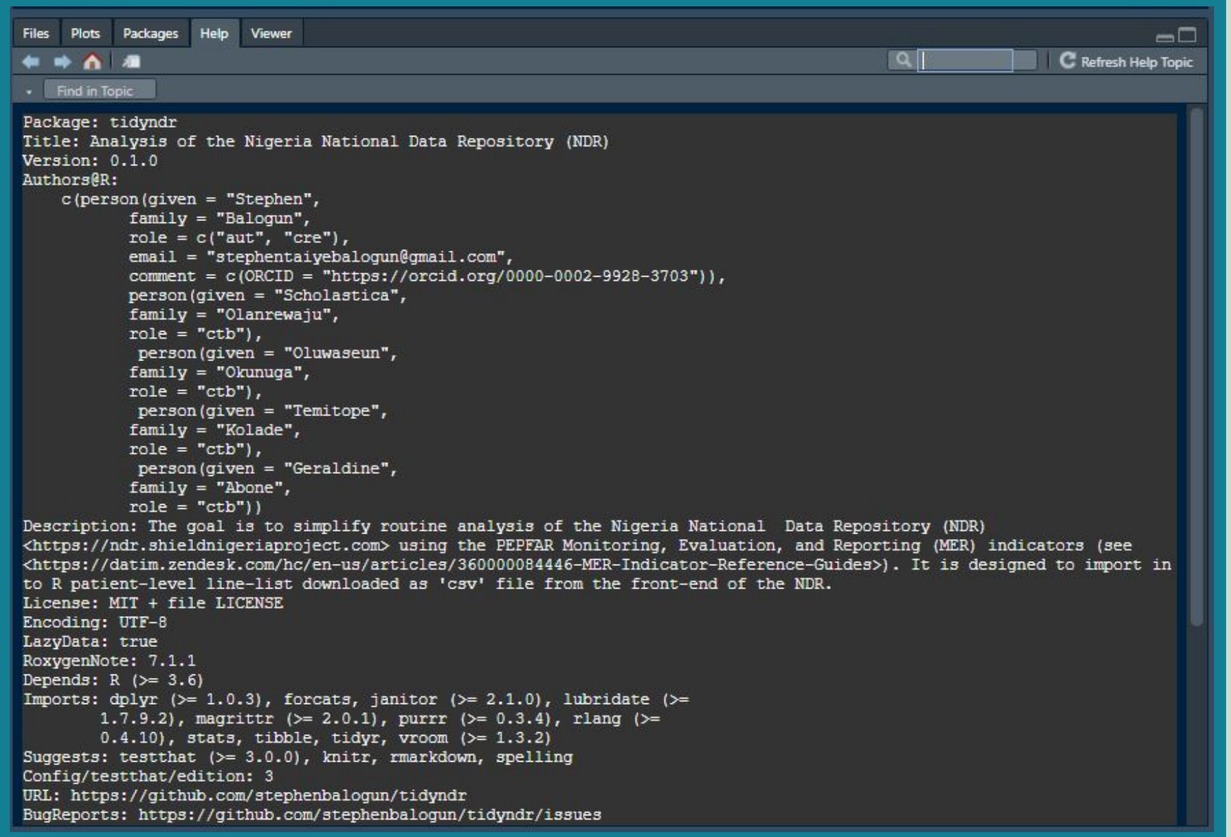
This analysis explores a data set containing 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others. The **goal** of the analysis was to *assess factors that affect borrower's APR for the loans.*

ANALYSIS TOOLS

CONCLUSION

- Prosper Score, Credit Score, and Loan amount are negatively correlated to Borrower's APR
- Loan amount and prosper score, credit score and monthly income are positively correlated to loan amount.
- Borrowers that are home owners have lower borrower APR compared to non-home owners. Furthermore, home owners have higher credit score and access to higher loan amount.
- Although loan amount and term has a positive relationship, on further exploration with a multivariate plot, it came out that when Borrower's APR is included it neutralises the effect of term on loan amount.

Develop an R Package- tidyndr



BACKGROUND

The Nigeria National Data Repository (NDR) houses the de-identified patient-level information for the HIV program in Nigeria. It allows users with login access to download deidentified patient-level data.
However, the analysis of this routine data proved challenging to many users due to the complexity of the algorithm used to define the indicators. Also, the size of the data make it difficult to use traditional spreadsheet tools for this analysis.

tidyndr was developed with the goal *is to "simplify routine analysis of the Nigeria National Data Repository (NDR)"*. It is designed to import in to R patient-level data downloaded as 'CSV file from the front-end of the NDR. With its in-built functions and a few lines of codes, it generates an analysis of the key performance indicators.

tidyndr was presented at the useR 2021 conference. tidndr is available for download from CRAN.

MY ROLE

I was a **contributor** on the tidyndr projects. I reviewed the codes and functions to validate that it is accurate. I tested the functions to validate that the output is as expected when compared against the traditional method of analysis.

Let's Work Together

s.olanrewaju.17@aberdeen.ac.uk

+234 908 304 1000

@Solanrewaju

@Solanrewaju