**Data Science Project**

**Contents**

**Section Page**

Executive Summary 2

Data Sources, Processed and Preparation 2

Analysis Performed 3

Visualisations 4 – 6

Recommendations for future iterations and possible improvements to the project 6 – 7

References 7 – 8

* + - **Executive Summary.**

How can Visualizations and Dashboards be used to display game data from Roblox *(Kang, 2024)* to give KPI’s for both users and administrators to give the users options to consider in terms of their game play and administrators a greater understanding of trends.

* + - **Data Sources, Processed and Preparation.**

Roblox is an online game platform, often played on mobile devices or laptops. A user can create an Avatar which can be used to play in any of the large library of games stored on the platform. New games can be created by users themselves as well as dedicated studios. Below is the ETL pipeline diagram displaying how the Roblox data was sourced.

A diagram of a computer program

Description automatically generated

1. Roblox data is stored on the Web.
2. Python code was used to scrape the necessary data from the Web *(Thomas, 2019)* which was then held in Kaggle *(McCain, 2022)* and made available for anyone to view. The original Python code is also available on the site.
3. The Kaggle table was stored in a .csv format, this was downloaded to a local laptop drive.
4. Transformations were required to the data, Excel was used to do this as .csv files naturally open in Excel, the data size of 119k was not excessive and the tasks could be performed with relative ease. Below is an example of the data:

A screen shot of a computer

Description automatically generated

Two columns were inserted between the ‘Date’ and ‘Active User’ columns, the first of these new columns was titled ‘Date’ the second ‘Time’. The Date formula was used to get just the date element from the original ‘Date’, then the Time formula was used to get8 just the time element. This allowed reviews to be done by both date and time and also ensured appropriate formats for both elements. The formulas were removed by using Copy and Paste Special – Values, then the Original Date column was deleted.

The Dates in ‘Date Created’ and ‘Last Updated’ needed to be reformatted as they were held in the American style and needed to be in the UK style.

In the ‘Total Visits’ Column the figures are abbreviated with letters, two columns were added between ‘Total Visits’ and ‘Date Created’. Firstly the 2 characters to the right of the ‘Total Visits’ field were pulled into the first additional column, then Find and Replace was used to remove ‘K+’, ‘M+’ and ‘B+’ from the ‘Total Visits’ column. Excel then amended the figures remaining to numbers. A formula was used to multiple the numbers left in the ‘Total Visits’ by 1,000 if ‘K+’, a million if ‘M+’ or a billion if ‘B+’.

1. The formatted data was then uploaded to Power BI for analysis and display.
   * + **Analysis performed.**

Two considerations for analysis were the length of time between the Date and when the game was created and the time between the Date and when the game was last updated. To achieve this the number of days between the two dates and the ‘Date’ needed to be calculated, this was done as additional columns in the transform data step. The formula ‘Datediff’ was considered, however using ‘Date’ – ‘Date Created’ will give the days between the two dates *(Stack Overflow, 2019)*, the same formula was used to create the Days since Updated.

Options considered for use in the display were:

* Total Visits
* Total Active Users
* Total Visits and/or Active Users by Date
* Total Visits and/or Active Users by days between Date and Date Created
* Total Visits and/or Active Users by days between Date and Date Updated
* Total Visits and/or Active Users by Genre
* Total Visits and/or Active Users by Title
* Total Visits and/or Active Users by Creator
* Total Visits and/or Active Users by Category
  + - **Visualisations.**

Below is a Screenshot of the final dashboard.

A screenshot of a computer

Description automatically generated

The Sum of Total Visits was favoured over Active Users as the Total Visits showed the success of the game whereas Active Users is better for a snapshot view at a given time.

A Time Series Analysis *(Velicer, 2003)* was used on Active Users to show the increasing use of the Game and potential date trends. Total Visits was not used for this as the constant increase would be difficult to see if patterns were increasing or decreasing.

The use of a Combined Line and Column graph for both Total Visits and Active User by Genre was to demonstrate the most popular Genre overall as well as the most popular at a particular time. Category was also considered for this type of graph, however the Category column only had three possible values and the majority of records fell into the same Category making this graph not very useful.

Title was used next, this relates to the name of the Game in Roblox, there are far too many to view all so a bar graph was used, and a filter applied to select the top 5 *(Russo, 2020)*. Separate graphs were used for Total Visits and Active Users as this shows which have been the most popular games of all time and what are the most popular games at the time.

A screenshot of a computer

Description automatically generated

Creator was used in a similar way to Title to again demonstrate the all time most popular creators and the current popular creators.

The final charts used Days Since Created and Days Since Updated. This time only Active Users were used, not Total Visits as more time will, inevitably, lead to more visits so Active Users would be a stronger metric.

In order to create the Columns the Days between factors were grouped using automatic grouping *(Microsoft, 2024)*, an example of how this was done is below.

A screenshot of a computer

Description automatically generated

On the side are two slicers, one for Genre and one for Creator. This is to enable the user to examine the data split by these factors or even a combination of both factors.

* + - **Recommendations for future iterations and possible improvements to the project.**

As the original Python code is included this could be adjusted to use a larger time frame and more up to date data. This would allow for the Time Series analysis to have more power as it could start to see potential seasonal trends.

Larger data would mean that the transformation steps performed by Excel would become difficult if not *impossible, this could be moved into the original Python code (Oliveira, 2023) by setting the formats to* the necessary types, this would also be quicker than taking the extra steps.

The Python code could be automated *(Wheeler, 2021)* to run each month to continuously update this report. May be best to use a rolling two years’ worth of data, this should be enough to establish trends over time but without using some much data as to confuse the message and slow down Power BI.

An improvement could be to add a type of slicer which would allow the user to select the number of Titles or Creators shown in the Top 5 charts *(Smoak, 2019)*. Effectively the slicer would change the 5 to a value selected by the user enabling them to see more information.

The Days between graphs have been shown as Column graphs, this shows the trend that the newer created games have more Active Users, and even stronger for more recently updated games. A scatter graph would also show this trend and may be better, especially if the data were larger *(Slutsky, 2014)*.

The section most likely for improvement is the use of the Description column, using trend analysis of the text *(Ke, 2024)* could open up a strong analysis into the links between the description and the popularity of the game. An extra section could be considered by allowing the user to select a number of their favourite games and then recommendations could be made by finding games with similar descriptions to the descriptions of those selected.

Final considerations go towards data bias. As there are no identifiable data in this data set there are no ethical considerations to be concerned about. However, if this Dashboard were to be used much wider, then the resulting trends could become self-fulfilling as the games and creators suggested on the dashboard could then become more popular.

* **References**

Kang, 2024. *Entertainment Computing.* [online] Available at <https://www.sciencedirect.com/science/article/abs/pii/S187595212400065X?dgcid=rss_sd_all> [Accessed 23rd August 2024]

Thomas, 2019. *Data Analysis by Web Scraping using Python.* [online] Available at <https://www.researchgate.net/publication/335576922_Data_Analysis_by_Web_Scraping_using_Python> [Accessed 23rd August 2024]

McCain, 2022. *Roblox Games Data.* [online] Available at <https://www.kaggle.com/datasets/databitio/roblox-games-data> [Accessed 11th July 2024]

Stack Overflow, 2019. *what's the difference between DATEDIFF and doing subtraction by myself.* [online] Available at <https://stackoverflow.com/questions/55699829/whats-the-difference-between-datediff-and-doing-subtraction-by-myself> [Accessed 23rd August 2024]

Velicer, 2003. *Time Series Analysis.* [online] Available at <https://www.researchgate.net/publication/229633091_Time_Series_Analysis> [Accessed 23rd August 2024]

Russo, 2020. *Filtering the Top 3 products for each category in Power BI.* [online] Available at <https://www.sqlbi.com/articles/filtering-the-top-3-products-for-each-category-in-power-bi/> [Accessed 23rd August 2024]

Microsoft, 2024. *Use grouping and binning in Power BI Desktop.* [online] Available at <https://learn.microsoft.com/en-us/power-bi/create-reports/desktop-grouping-and-binning> [Accessed 23rd August 2024]

Oliveira, 2023. *A systematic literature review on the impact of formatting elements on code legibility.* [online] Available at <https://www.sciencedirect.com/science/article/pii/S0164121223001231> [Accessed 24th August 2024]

Wheeler, 2021. Creating *automated reports using python and Jupyter notebooks.* [online] Available at <https://andrewpwheeler.com/2021/07/10/creating-automated-reports-using-python-and-jupyter-notebooks/> [Accessed 24th August 2024]

Smoak, 2019*. Filter Top N Values with a Slicer in Power BI.* [online] Available at <https://www.youtube.com/watch?v=QtEt-QI3oe4> [Accessed 24th August 2024]

Slutsky, 2014. *The Effective Use of Graphs.* [online] Available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4078179/> [Accessed 24th August 2024]

Ke, 2024. *Recent Advances in Text Analysis.* [online] Available at <https://arxiv.org/abs/2401.00775> [Accessed 24th August 2024]