DPRPy 2022/2023

Homework Assignment# 3

By: Sushree Smaranika Pradhan

Dataset

In this assignment below data set has been used

1. Android.stackexchange.com.7z (link)
2. Aviation.stackexchange.com.7z (link)
3. Health.stackexchange.com.7z (link)

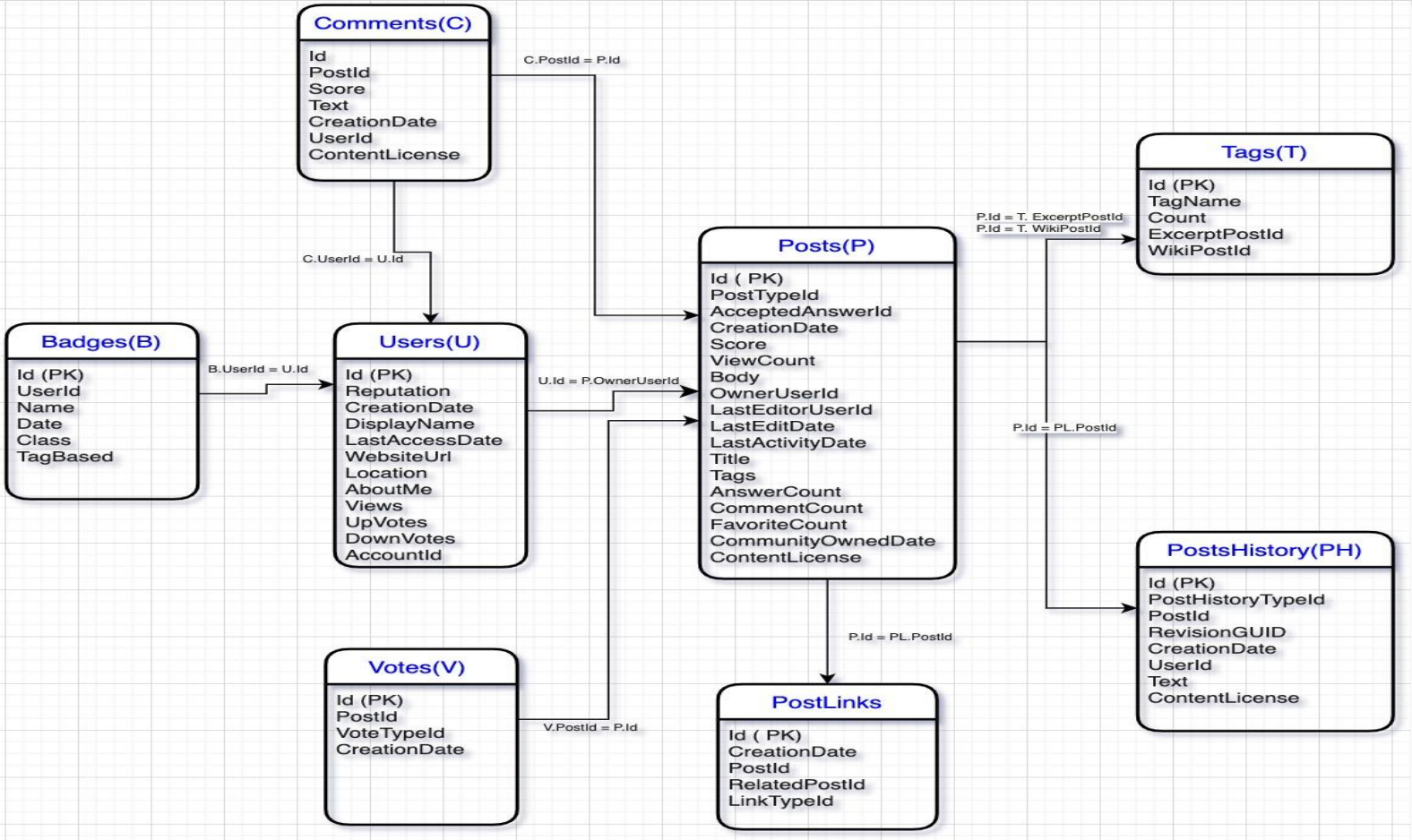
Note: All data files are downloaded from
https://archive.org/details/stackexchange

Dataset Description

This is an anonymized dump of all user-contributed content on the Stack Exchange network. Each site is formatted as a separate archive consisting of XML files zipped via 7-zip. Each site archive includes below data points

1. Posts
2. Users
3. Votes
4. Comments
5. PostHistory
6. PostLinks
7. Badges
8. Tags

Dataset Relationship Diagram



Libraries used

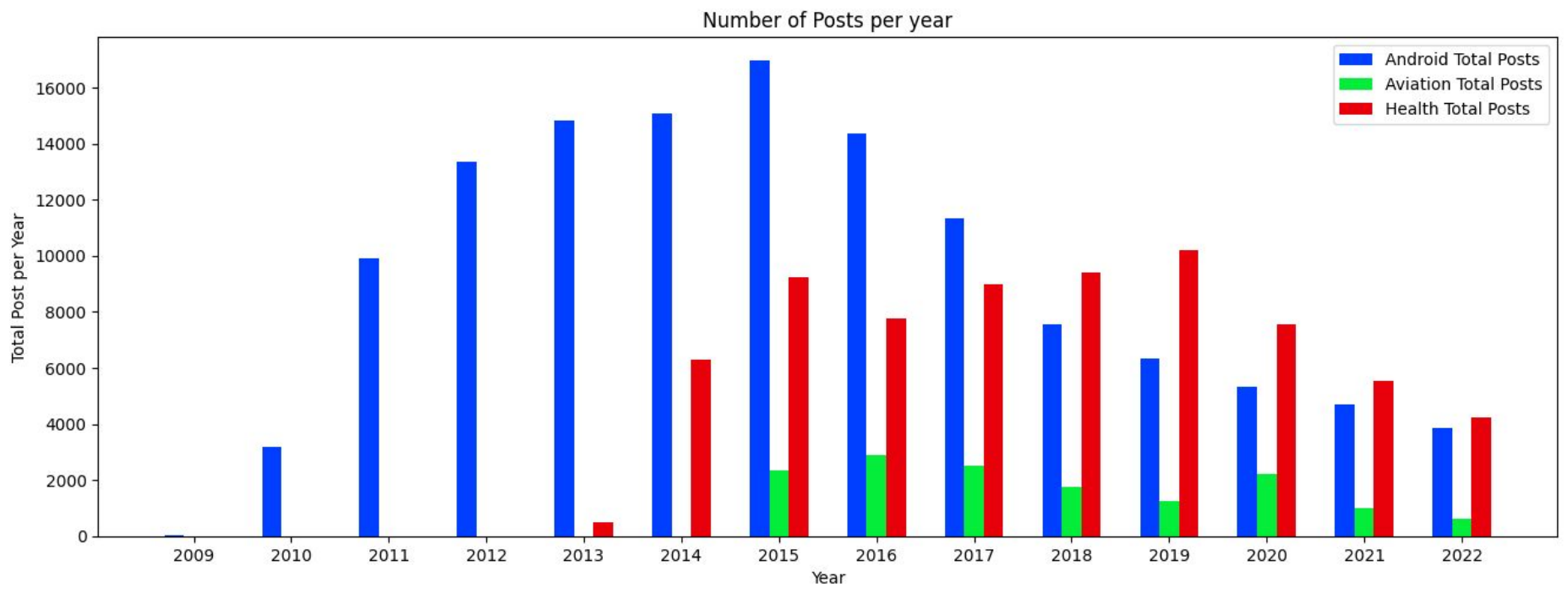
Library Name	Uses
numpy	Numerical Python
matplotlib.pyplot	Graphical Plotting library
xml.etree.ElementTree	Xml file parsing library
pandas	Python Data Analysis library
os	Operating System

User Defined Functions

Function	Uses
xml_2_dataframe	It parses the xml file and load the data into panda dataframe
read_xml_files	It will generate the a dictionary of dataframes from the xml files present with the data directory

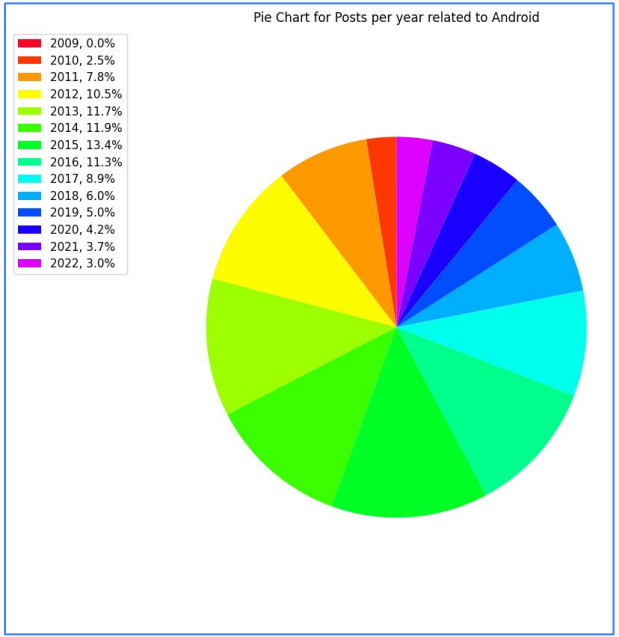
Data Analysis on Posts Datasets

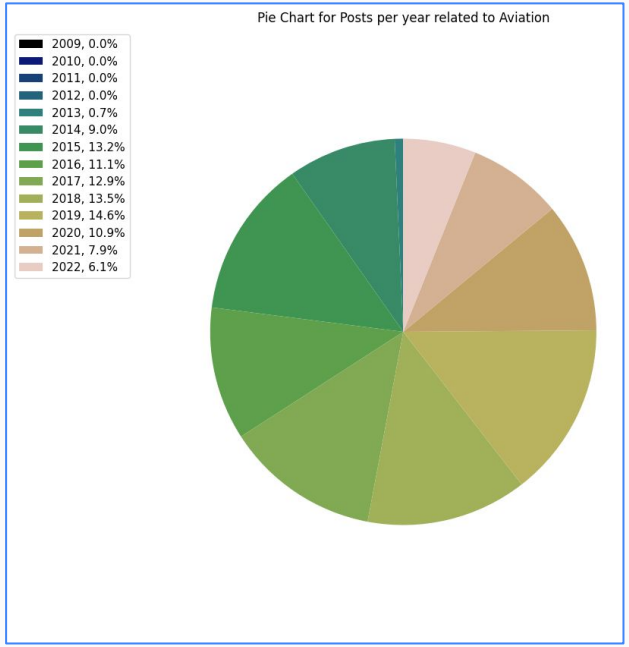
Goal: Identify the Number of Posts over the years
Conclusion: Most of the Posts created in 2015 for Android, 2019 for Health and 2016 for Aviation. And the overall trend the number of Posts are reducing over the years.

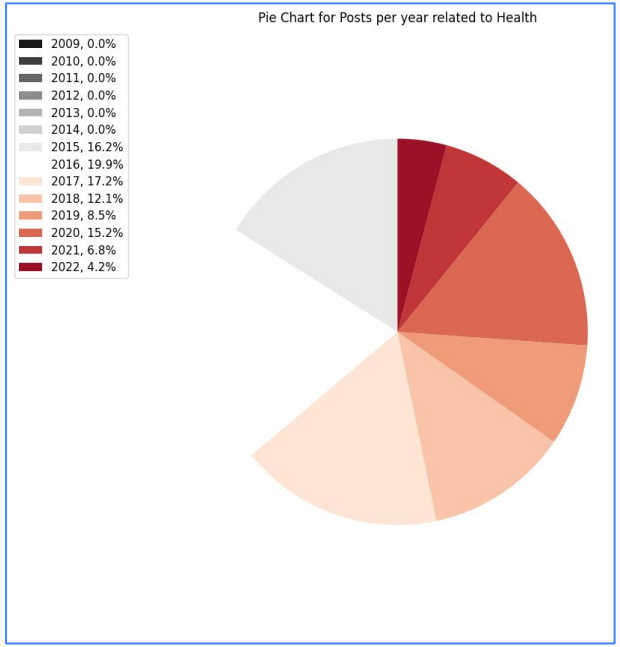


Data Analysis on Posts Datasets

Goal: Represent the dataset in pie chart

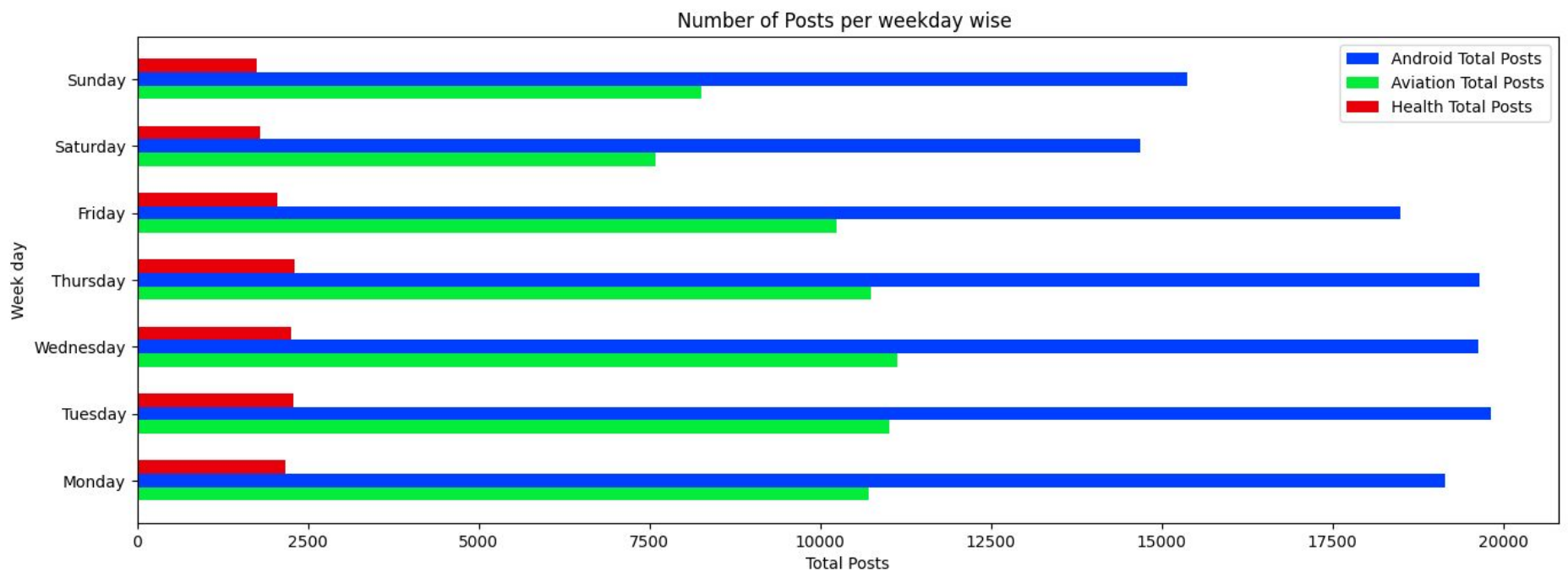






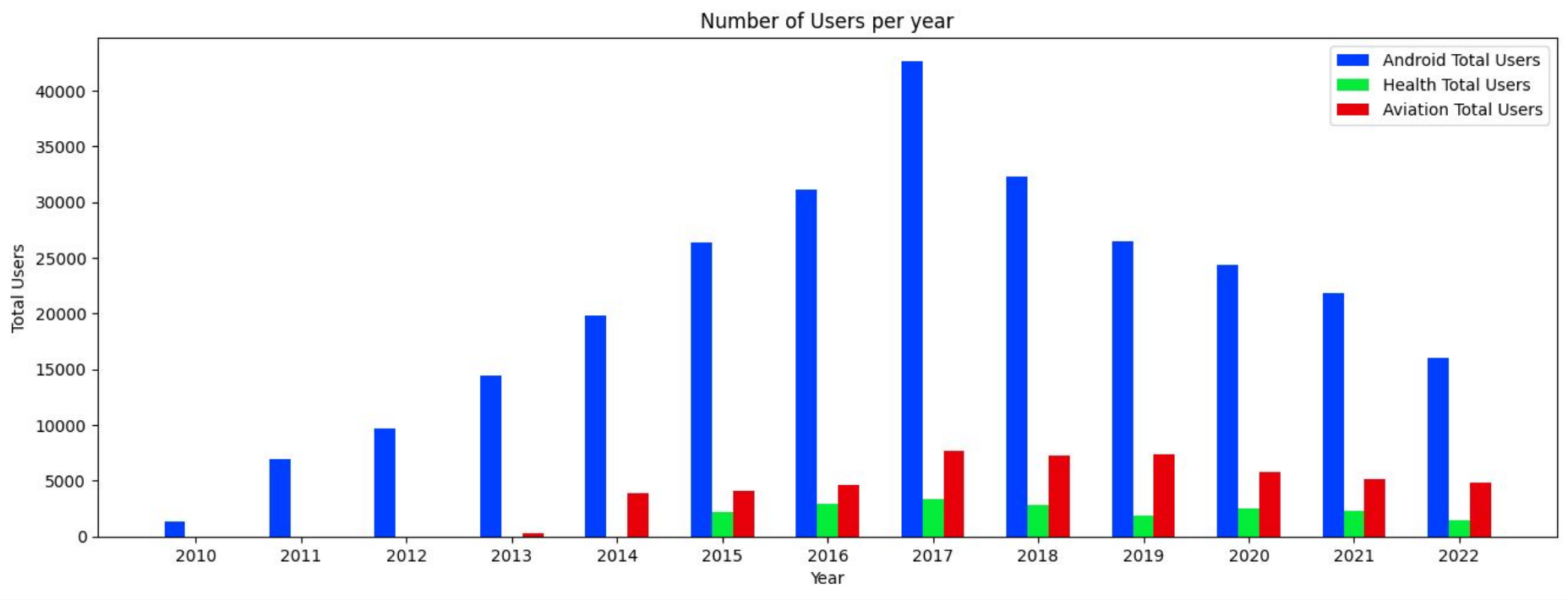
Data Analysis on Posts datasets

Goal: Identify the Number of Posts over weekdays
Conclusion: There are less number of Posts in weekend as compared to week days



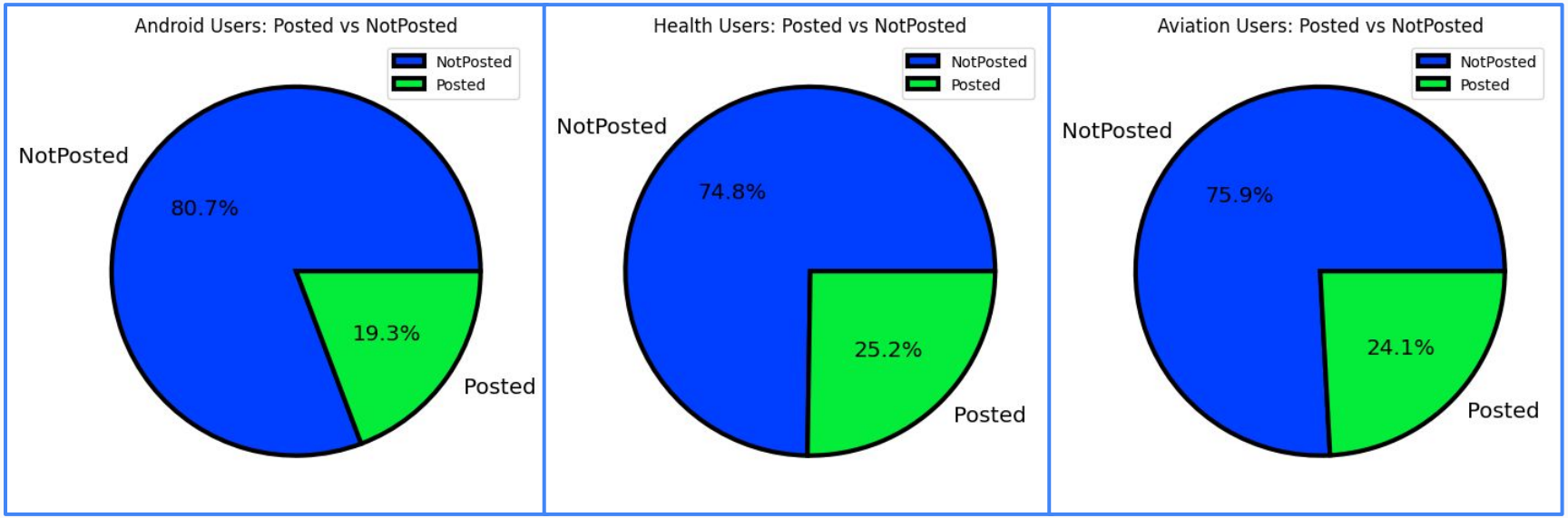
Data Analysis on Users datasets

Goal: Identify the Number of Users created over the Year
Conclusion: Most of the users created in 2017 for Android, Health and Aviation.



Data Analysis on Users datasets

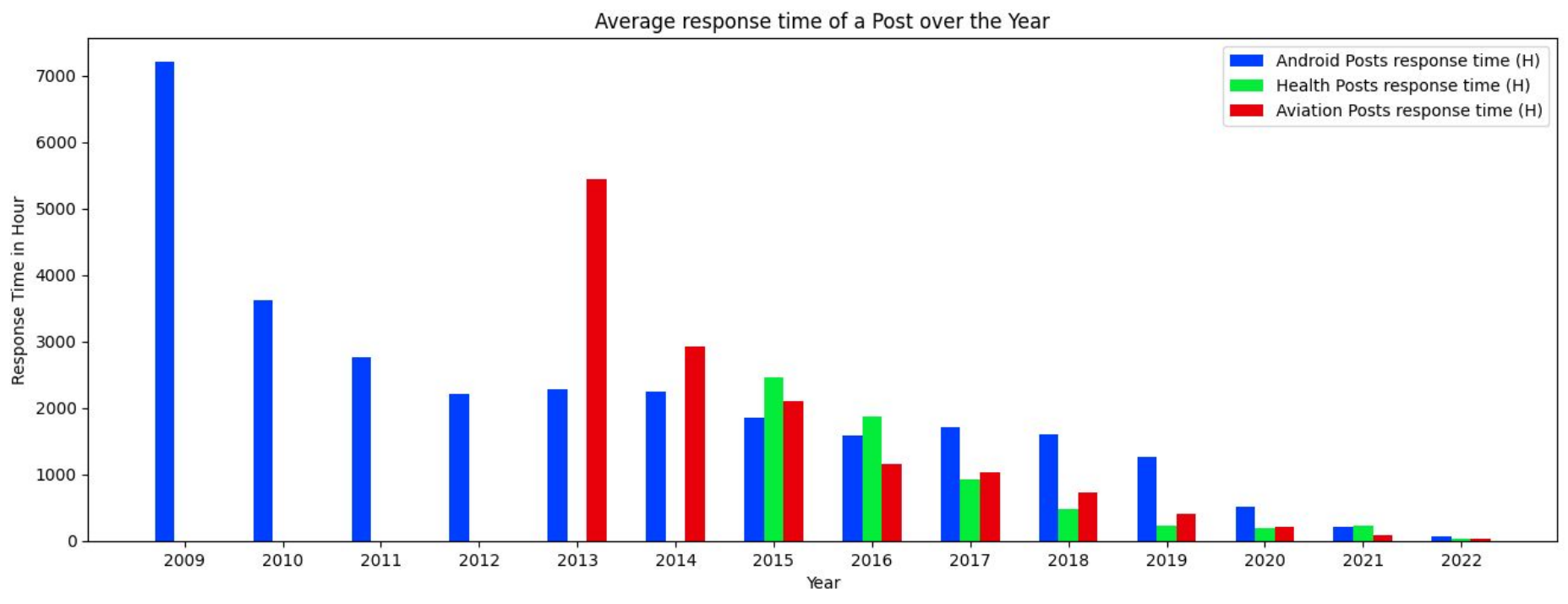
Goal: Pie chart for Number of users who Posted vs Not-Posted in the Portal



Data Analysis on Comments vs Posts

Goal: Identify the avg response time of a post over the year

Conclusion: From the analysis it has been observed that the response time of a post is reducing over the time.



Thank You !!