

CSC8631

Coursework

Data Management and
Exploratory Data
Analysis





“

“Non-reproducible single occurrences are of no significance to science”

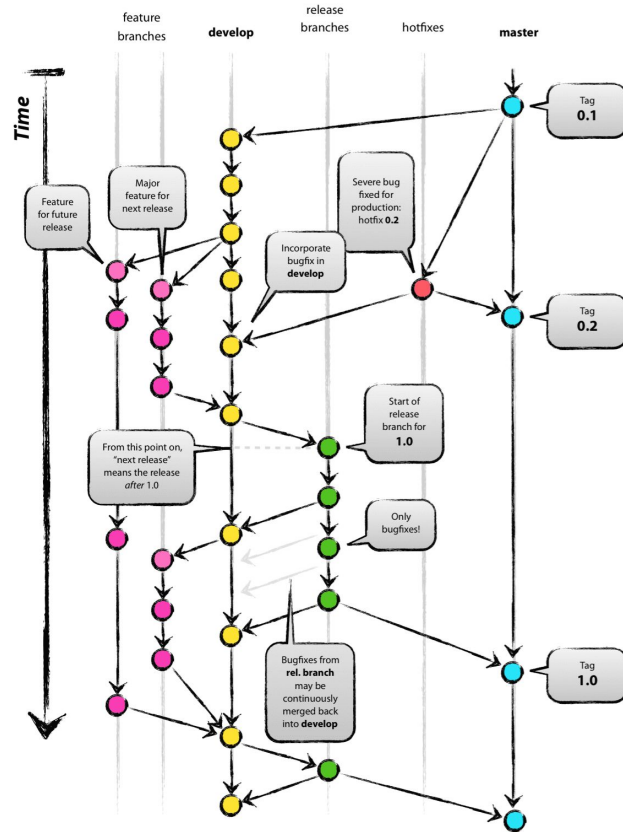
Popper (The logic of Scientific Discovery)

Git

Version Control System



Git Branching Model



Github Actions

Pull Request on Dev Branch

The screenshot shows the GitHub Actions page for the repository 'muzaffersenkai / CSC8631-Project'. The 'Actions' tab is selected, showing a list of workflow runs. On the left, under 'Workflows', there are two workflows: 'Deploy to Firebase' and 'Unit Test'. The 'All workflows' section shows two workflow runs: 'Deploy to Firebase' (status: success, branch: master) and 'First Release' (status: success, branch: dev).

Workflow	Status	Branch	Actor
Deploy to Firebase	Success	master	2 days ago
First Release	Success	dev	2 days ago

Merge on Master Branch



Agile Methodology - Kanban

The screenshot displays a GitHub Kanban board for the repository `muzaffersenkhal / CSC8631-Project`. The interface includes a top navigation bar with links for Pull requests, Issues, Marketplace, and Explore. Below this, a secondary navigation bar shows options like Code, Issues (1), Pull requests, Actions, Projects (1), Wiki, Security, Insights, and Settings. The main area is titled **TODO** and shows a board with three columns: **To do**, **In progress**, and **Done**. The **To do** column contains one card: **Data Preparation Iteration 3** (#12), added by muzaffersenkhal. The **In progress** column contains one card: **Data Understanding Iteration 3** (#12), opened by muzaffersenkhal. The **Done** column contains five cards: **Data Preparation Iteration 2** (#10), **Data Understanding Iteration 2** (#8), **Business Understanding** (#1), **Data Understanding** (#3), and **Data Preparation** (#6), all opened by muzaffersenkhal. A search bar labeled 'Filter cards' and buttons for '+ Add cards', 'Fullscreen', and 'Menu' are located above the columns. A '+ Add column' button is on the right side of the board.

Search or jump to... Pull requests Issues Marketplace Explore

muzaffersenkhal / CSC8631-Project Private Unwatch 1 Star 0 Fork 0

<> Code Issues 1 Pull requests Actions Projects 1 Wiki Security Insights Settings

TODO Updated 13 minutes ago Filter cards + Add cards Fullscreen Menu

1 To do + ...

- Data Preparation Iteration 3 #12 Added by muzaffersenkhal

1 In progress + ...

- Data Understanding Iteration 3 #12 opened by muzaffersenkhal

5 Done + ...

- Data Preparation Iteration 2 #10 opened by muzaffersenkhal
- Data Understanding Iteration 2 #8 opened by muzaffersenkhal
- Business Understanding #1 opened by muzaffersenkhal
- Data Understanding #3 opened by muzaffersenkhal
- Data Preparation #6 opened by muzaffersenkhal

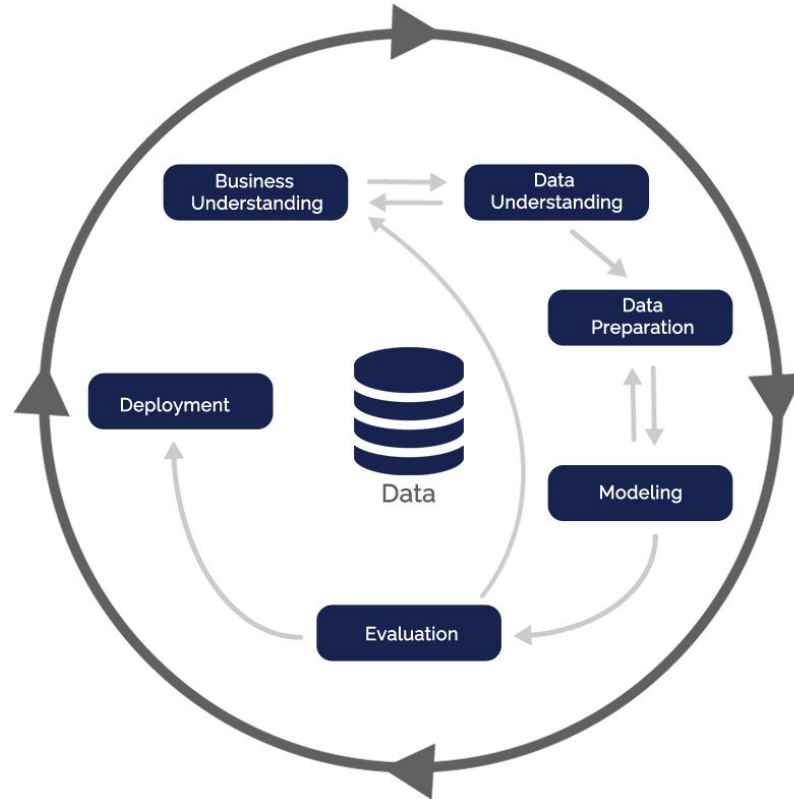
+ Add column

CRISP-DM

Methodology



Steps



Business Understanding



Estimated Reading Time

Businesses are making effort to achieve an edge with its content marketing program. Showing an article's reading time to each of article can have a profound and positive effect on reader engagement levels.





“

“I know that watching the film Pulp Fiction will take me exactly 154 minutes, and this doesn't change anything to the fact that it's an awesome film and that I will have a great time watching it. Knowing in advance how long an article will take me just helps me with my time management, by allowing me to plan better.”

Evidence

Business Objectives



**how the reading
time of the articles
changes according
to learners**



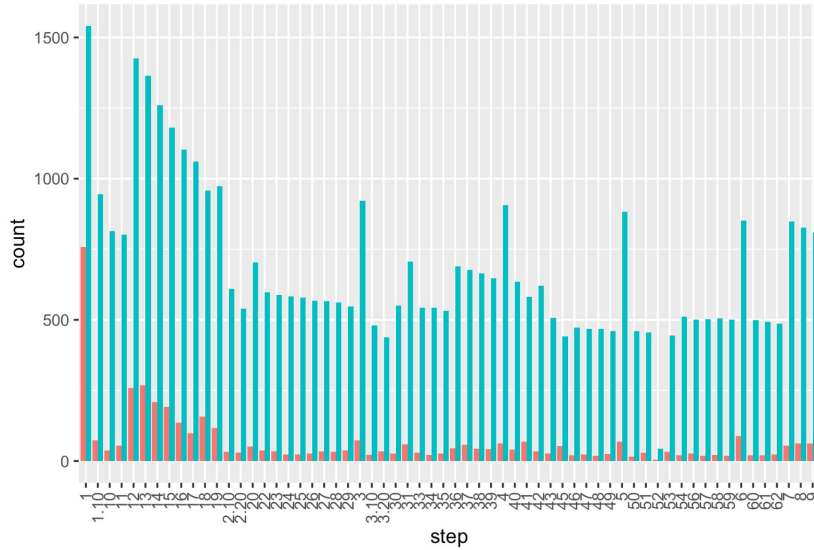
**Develop a model that can
predict the average article
reading time for each
person**

Exploratory Data Analysis

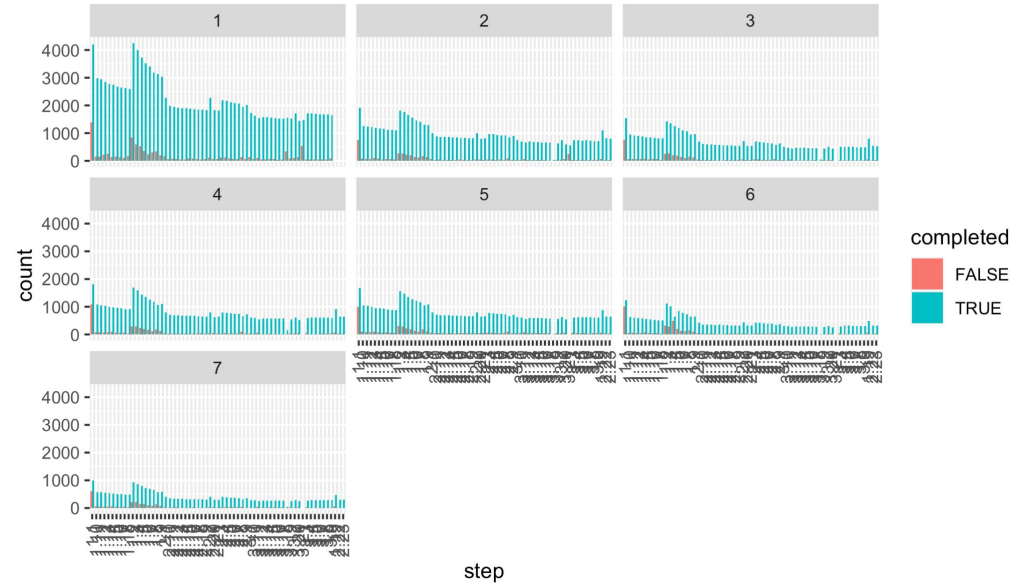


Step Activity

Run 3 - Number of Step Completed



All run - Number of Step Completed



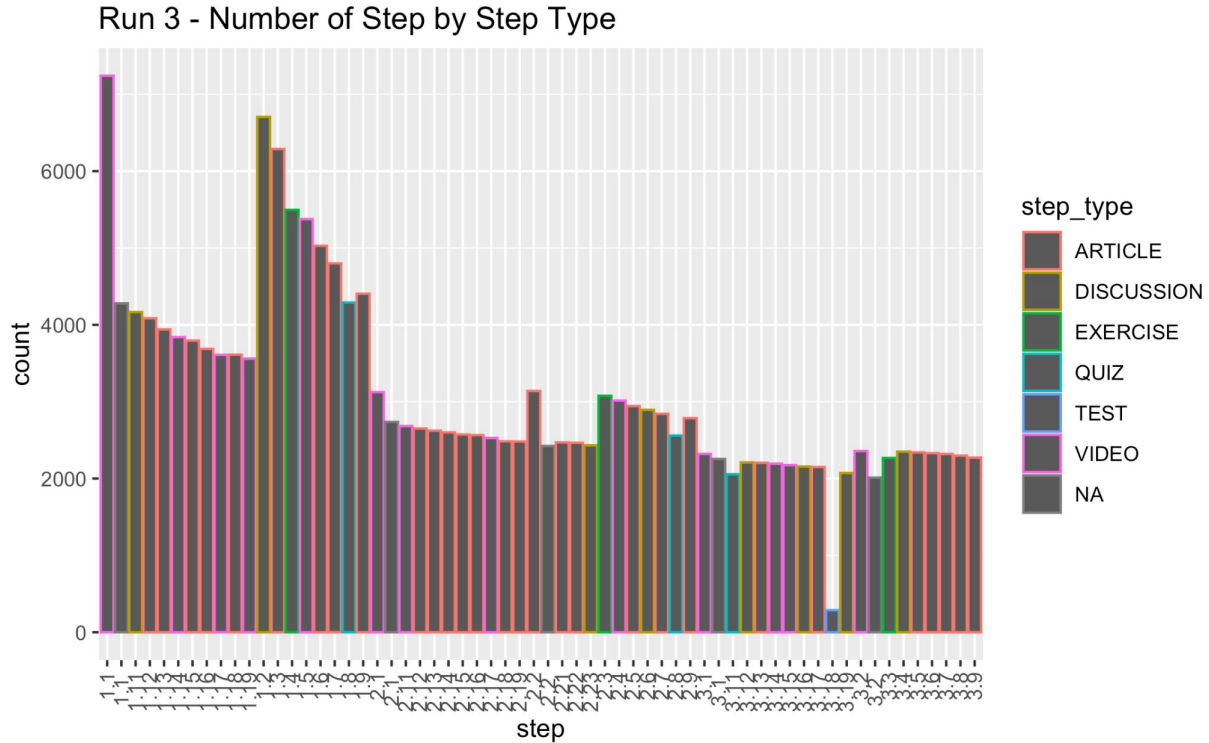
Step Activity

A tibble: 7 × 2

run_id <dbl>	n_distinct(step) <int>
1	60
2	63
3	62
4	62
5	62
6	62
7	62

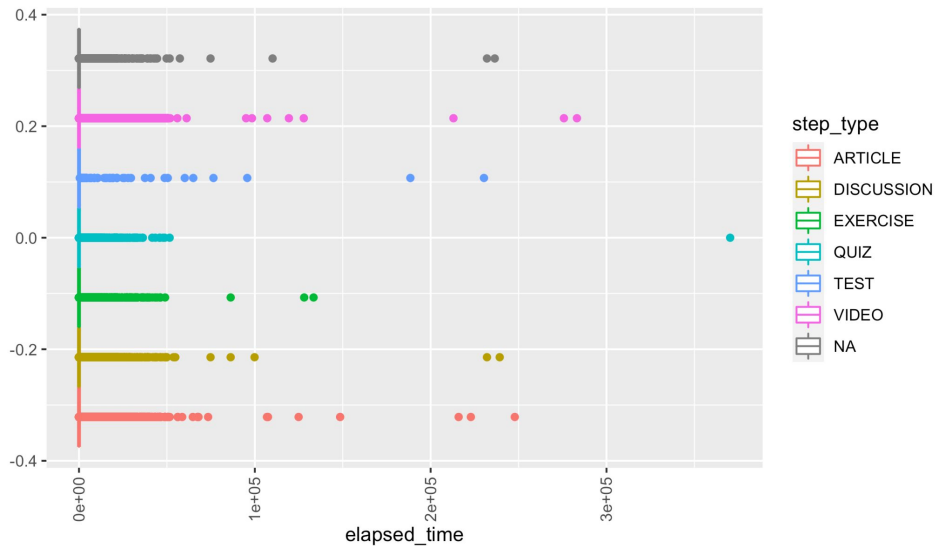
7 rows

Step Activity

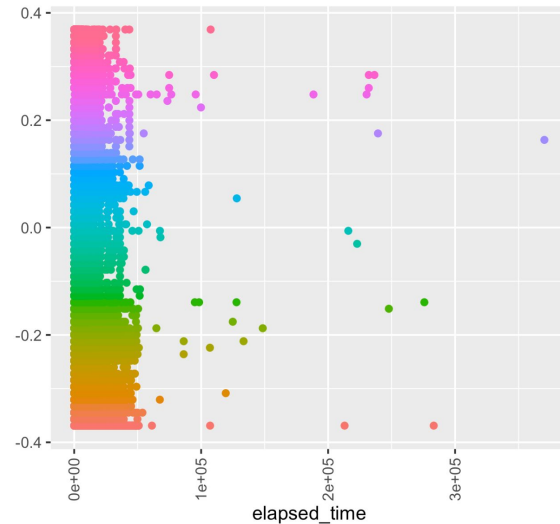


Reading Time

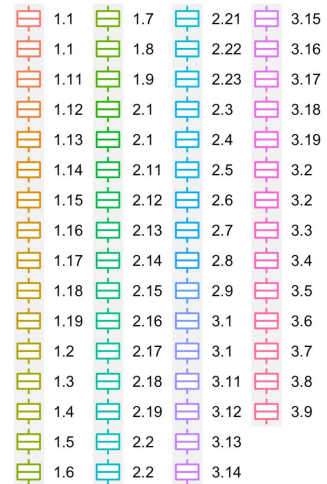
Boxplot of elapsed time for each type



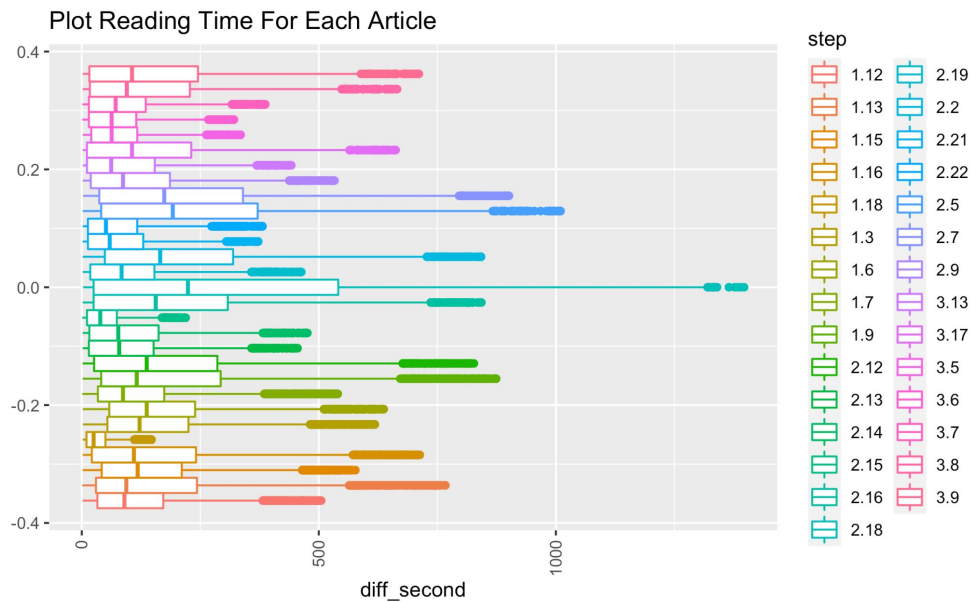
Elapsed time for each step



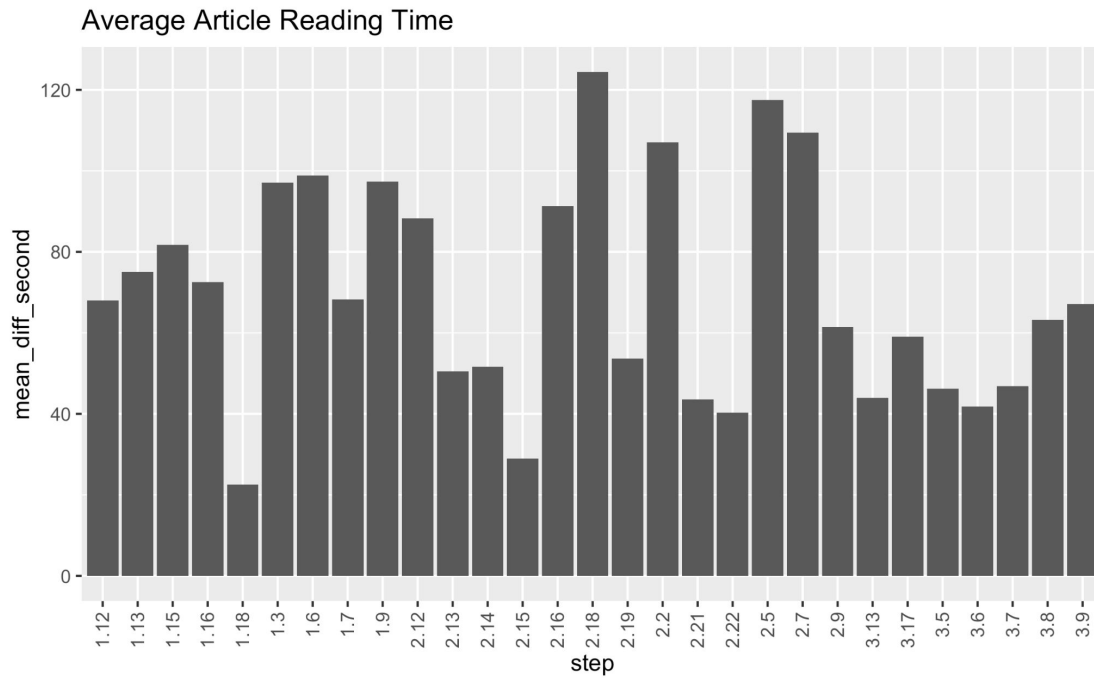
step



Reading Time

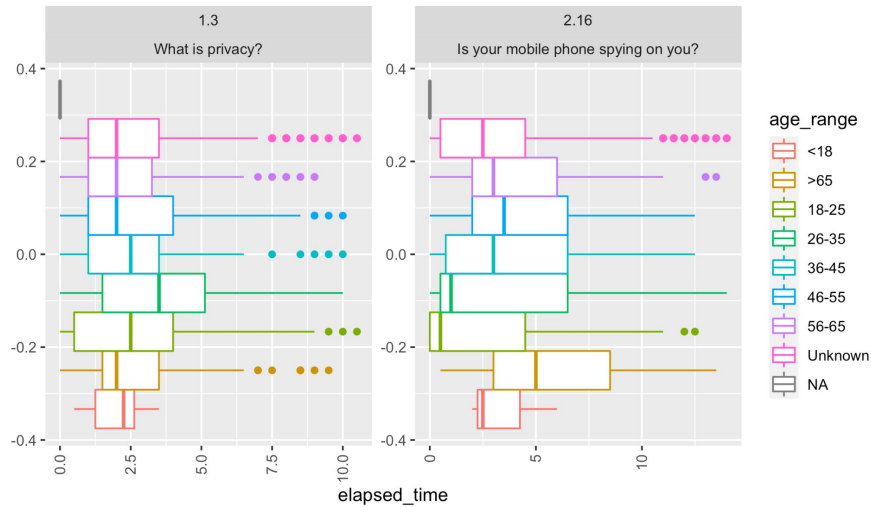


Average Article Reading Time

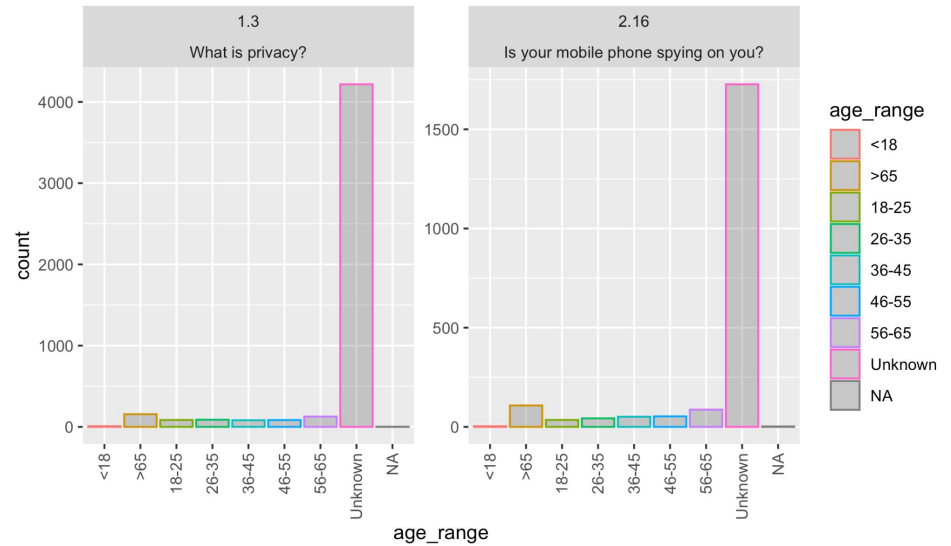


Age

Reading time of articles by age

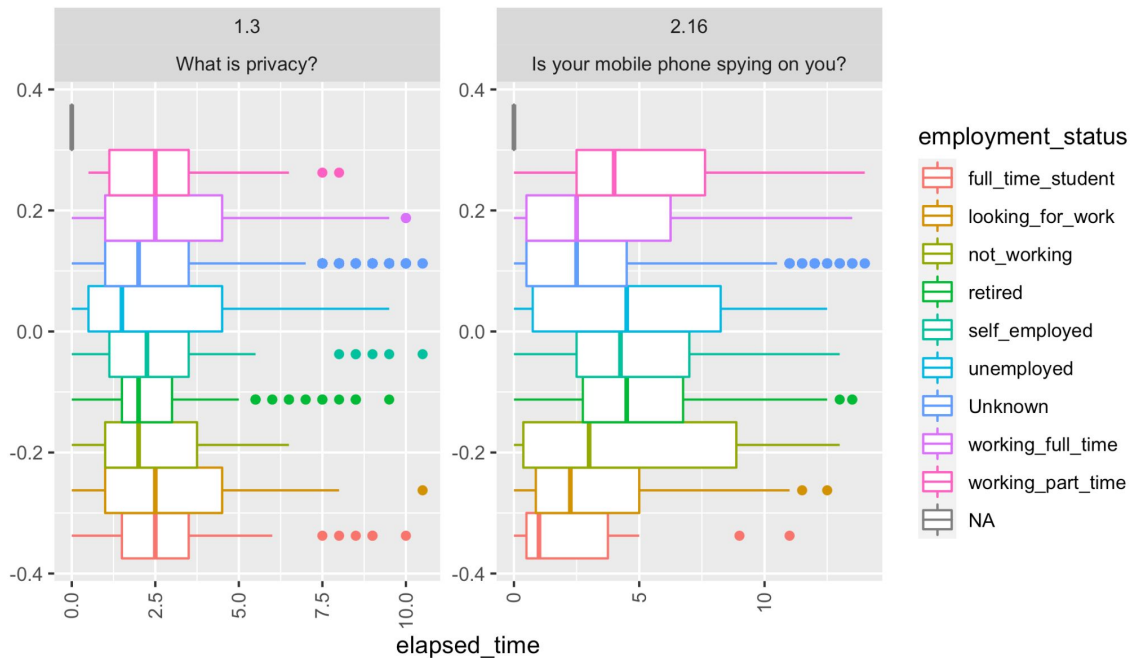


Age Distribution



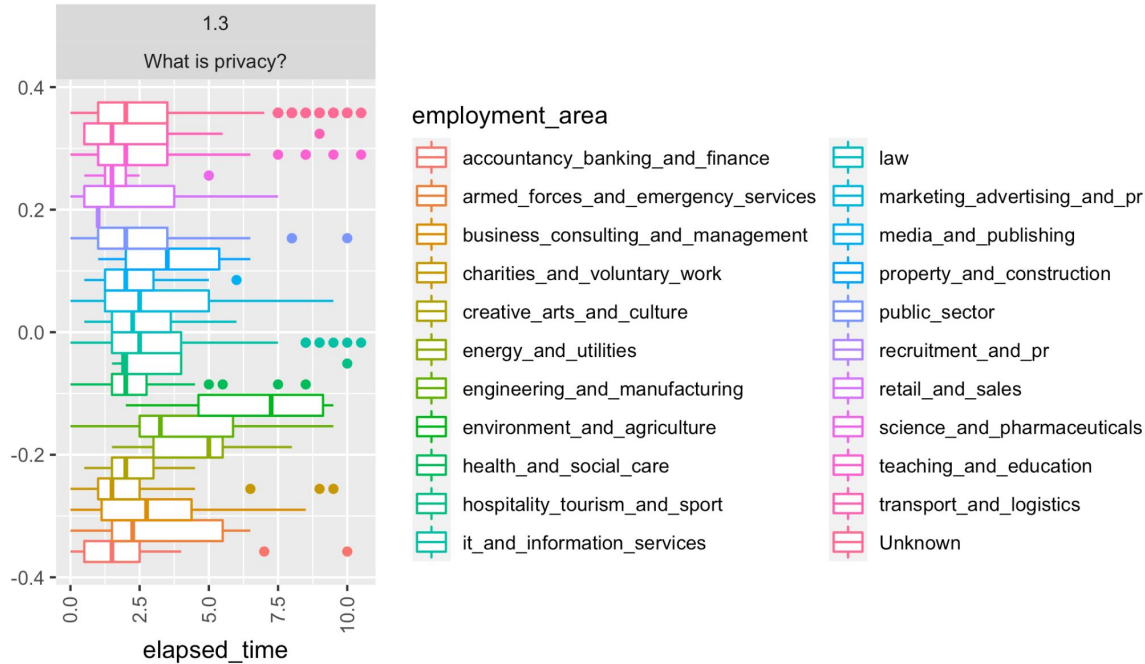
Employment Status

Elapsed Reading time of articles by employment status



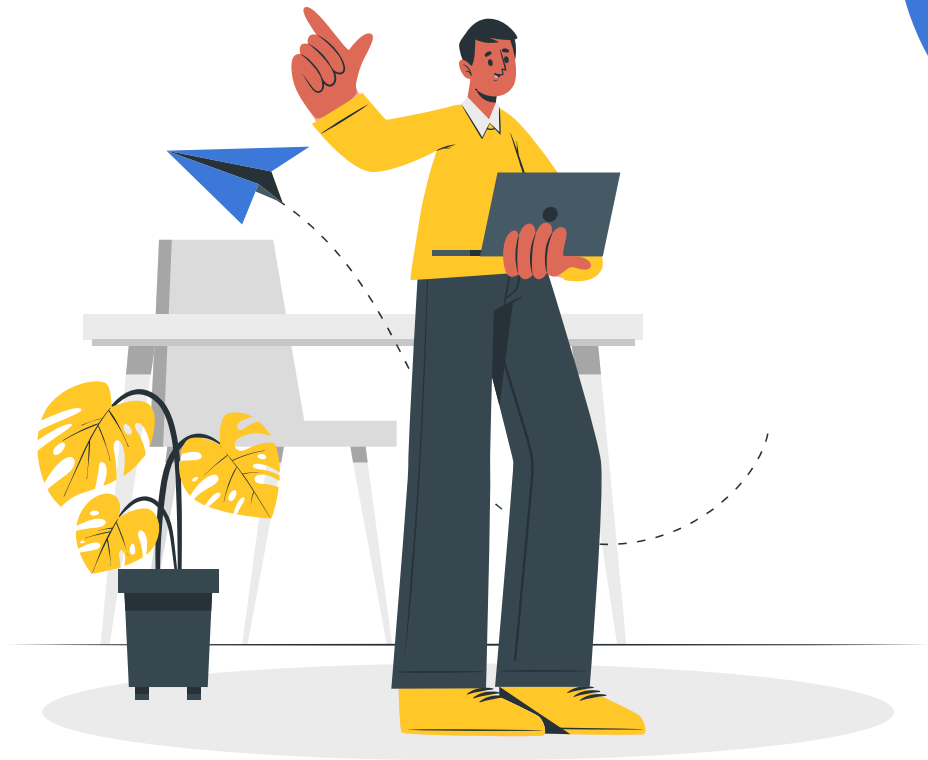
Employment Area

Elapsed Reading Time of Articles by Employment Area

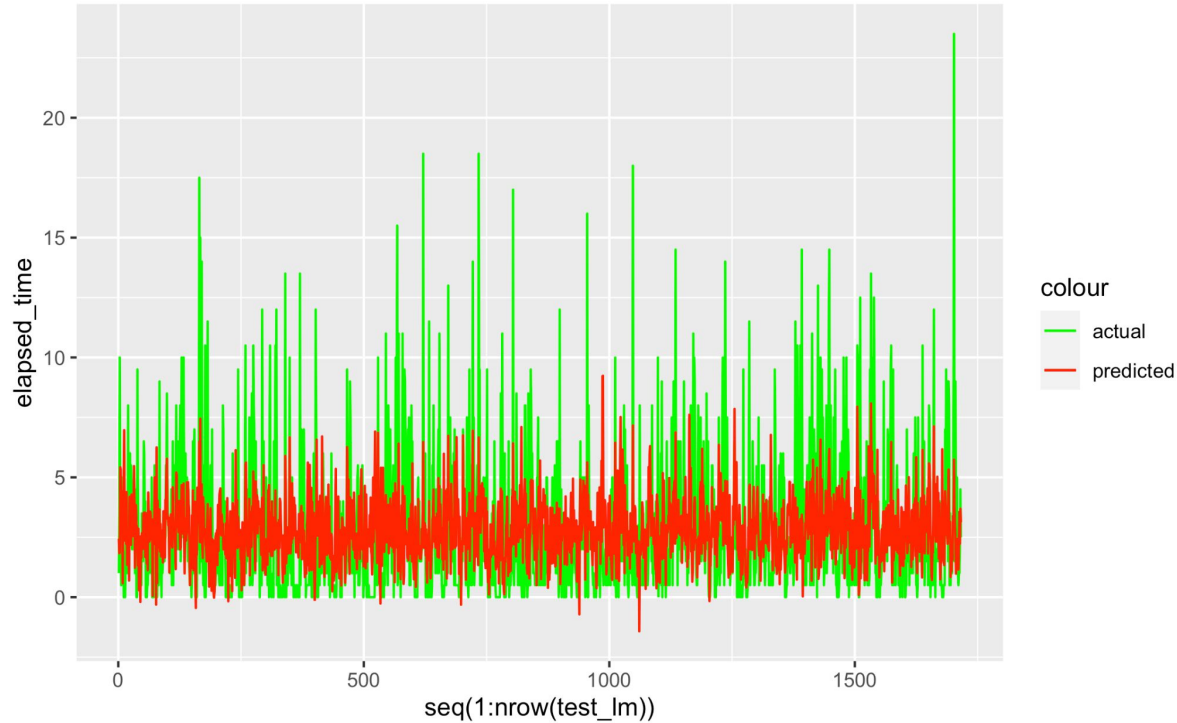


Model

Predict the estimated
reading time



Linear Regression



RMSE: 2.60 min

Linear Regression

