CSC8631 Coursework

Data Management and Exploratory Data Analysis

“

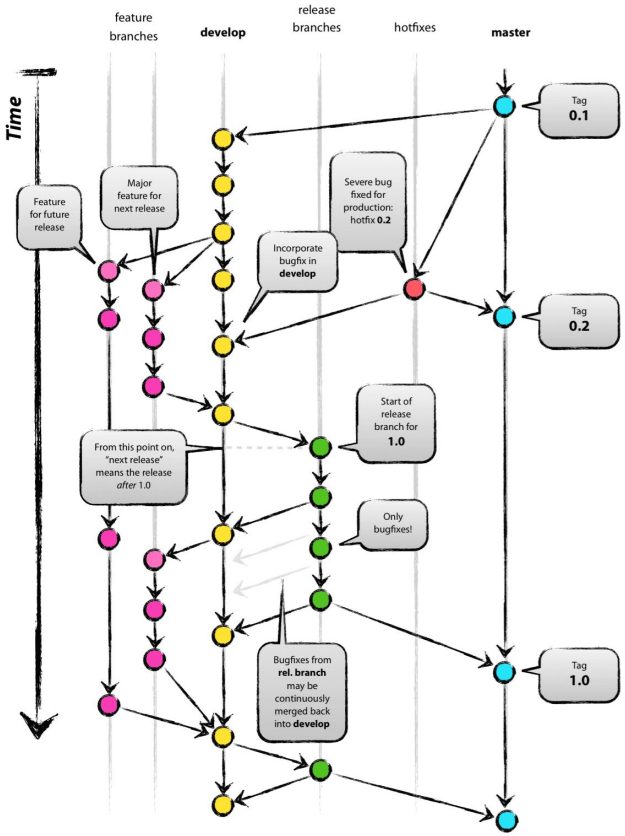
“Non-reproducible single occurrences are of no significance to science”

Popper (The logic of Scientific Discovery)

Git

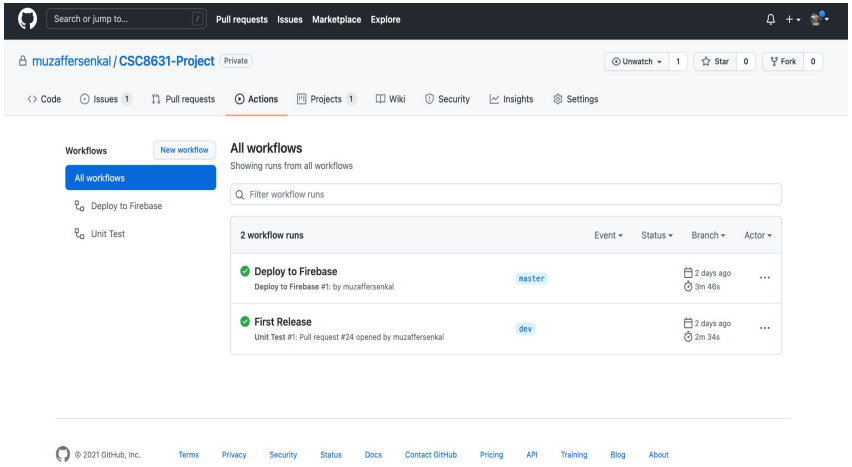
Version Control System

Git Branching Model



Github Actions

Pull Request on **Dev** Branch

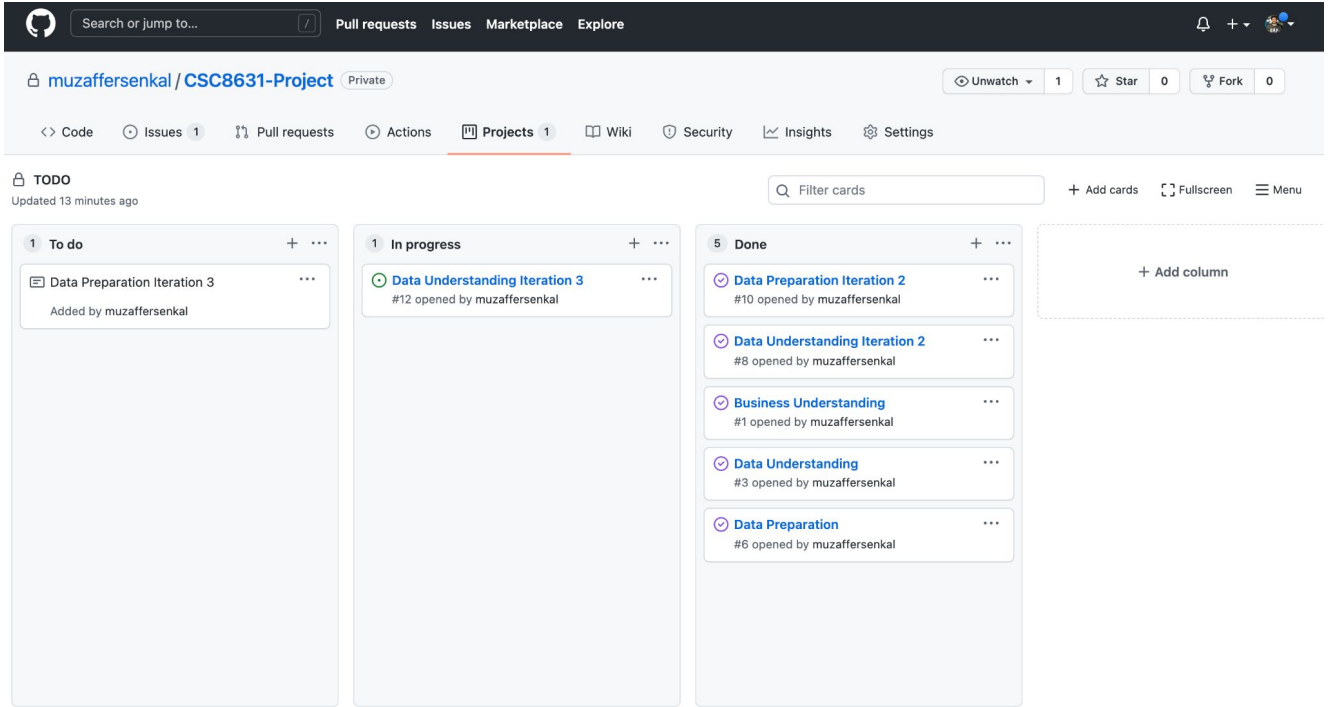


Merge on **Master** Branch

Firebase

https://datamanagement-4be96.web.app

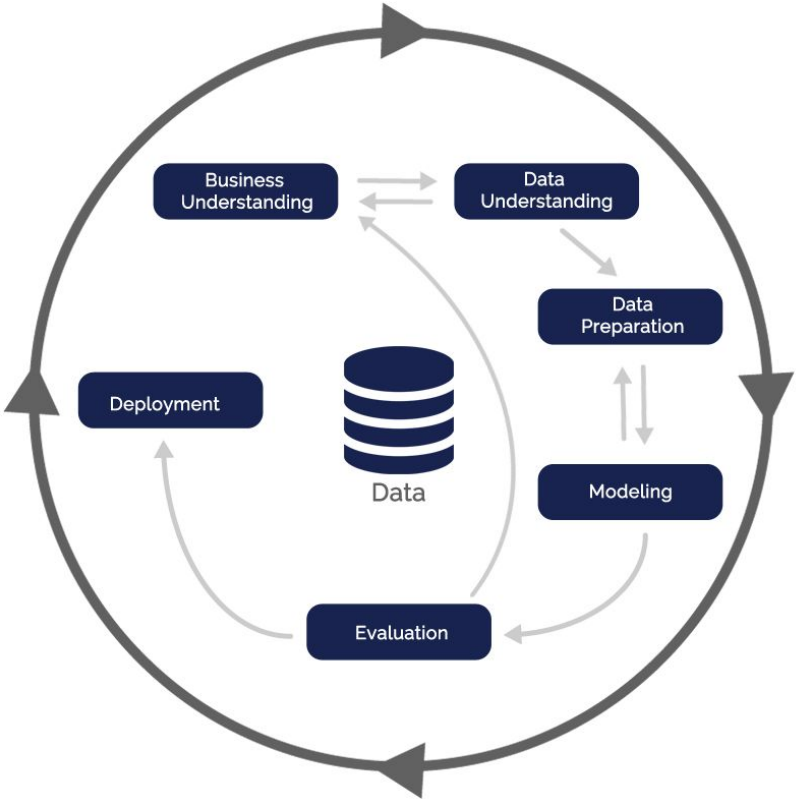
Agile Methodology - Kanban



CRISP-DM

Methodology

Steps


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

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

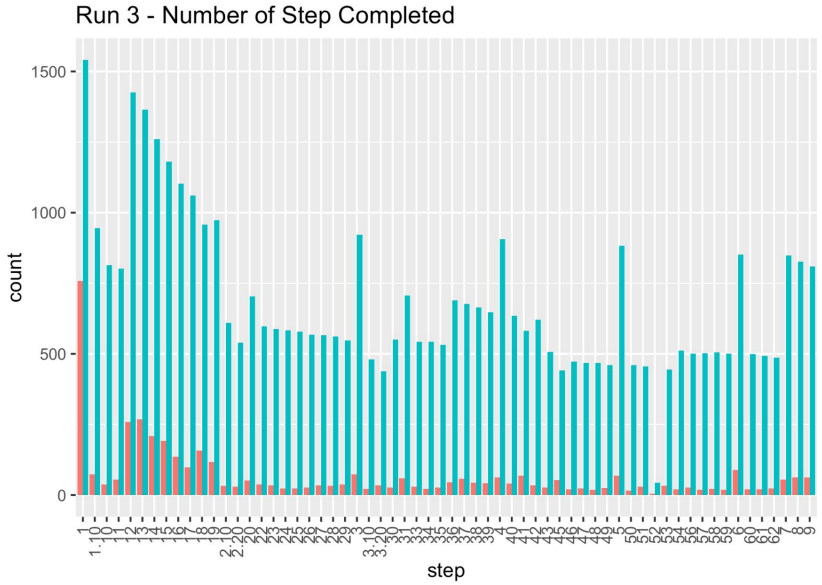
Business Objectives

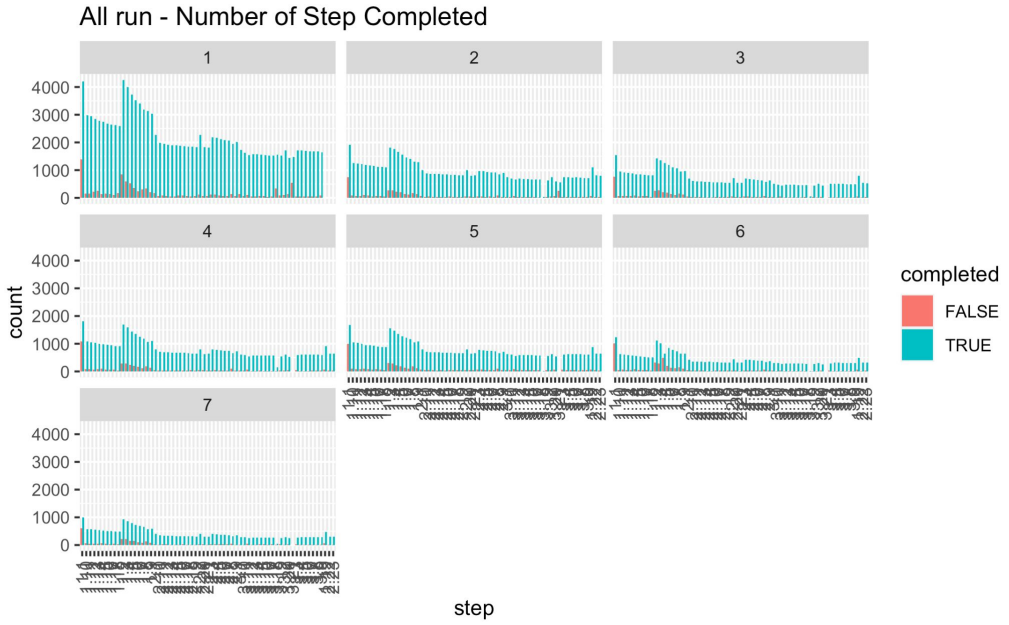
how the reading time of the articles changes according to learners

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

Exploratory Data Analysis

Step Activity





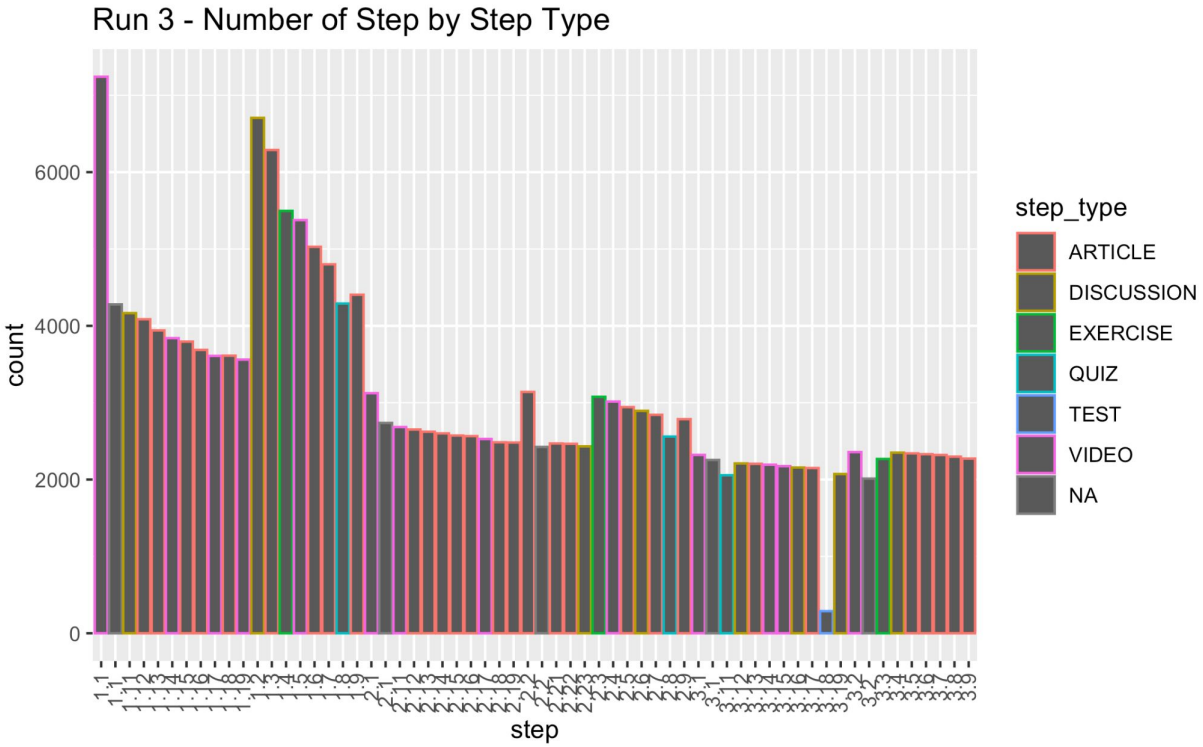
Step Activity

A tibble: 7 × 2

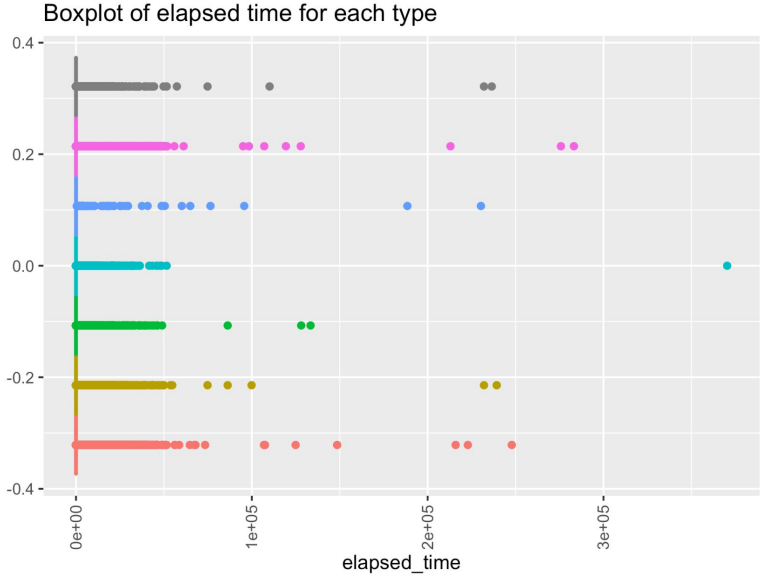
run_id <dbl>	n_distinct(step) <int>
1	60
2	63
3	62
4	62
5	62
6	62
7	62

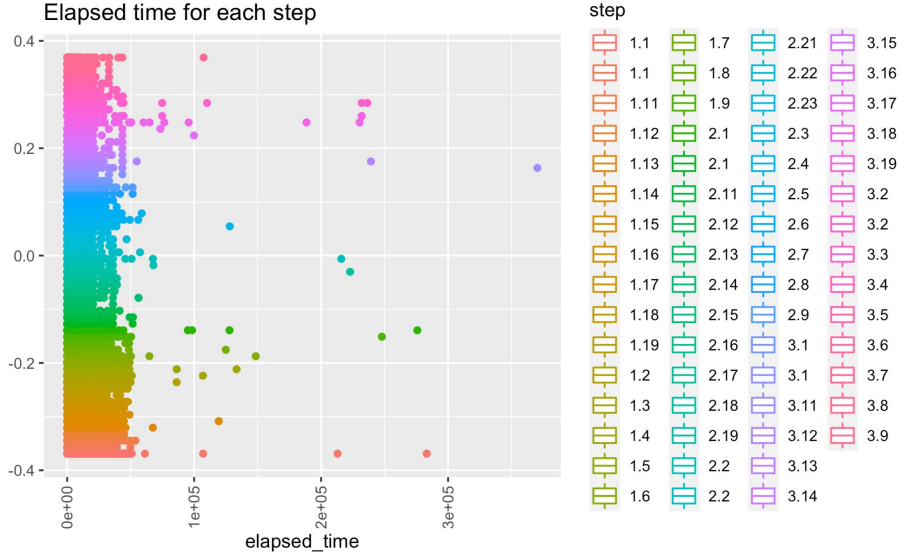
7 rows

Step Activity

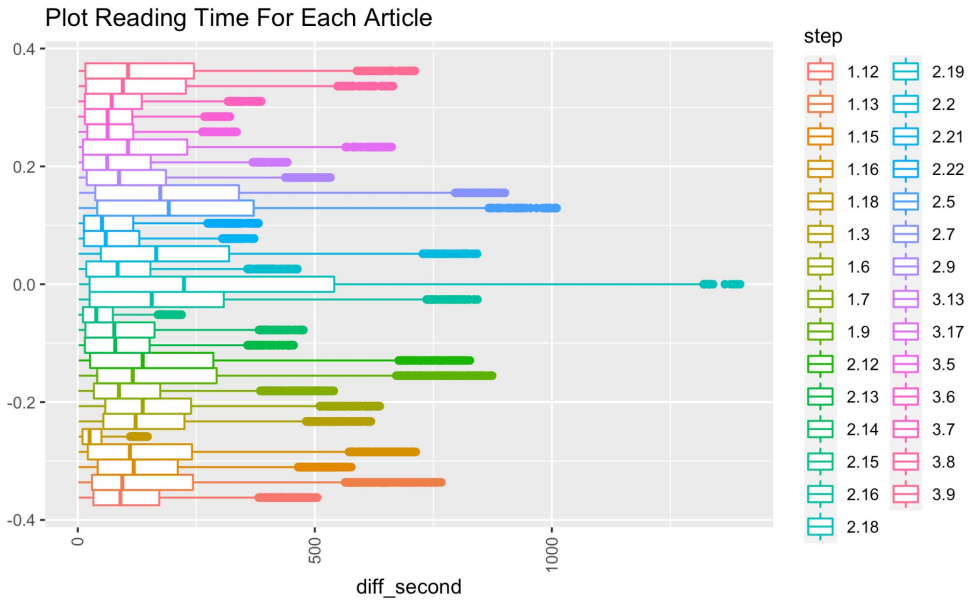


Reading Time

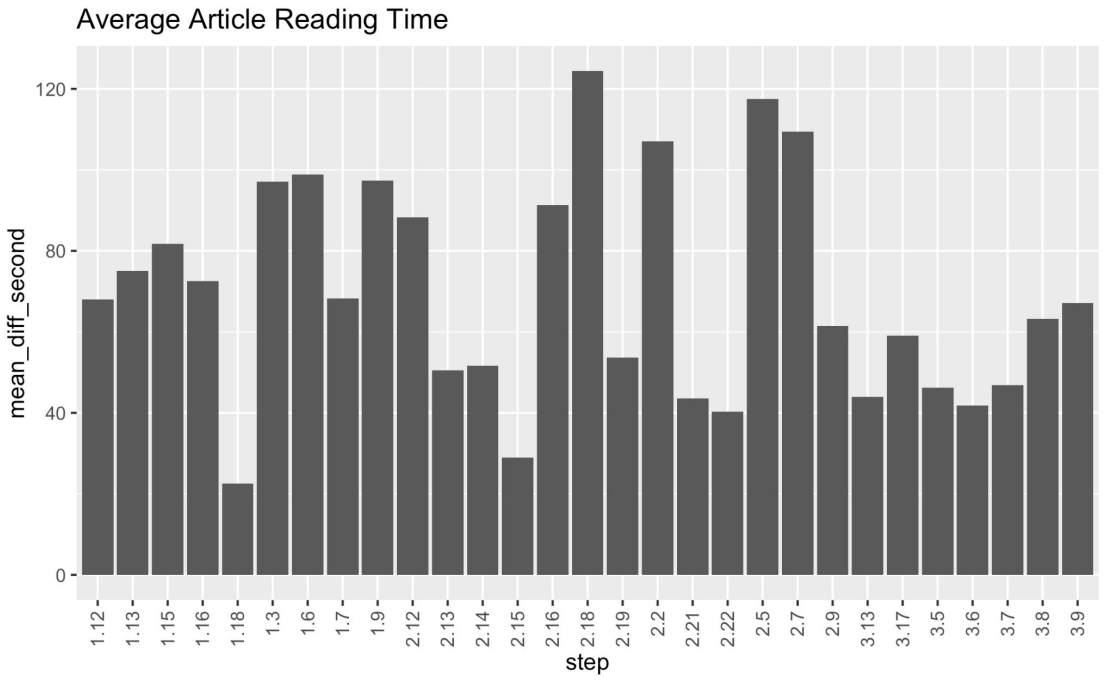




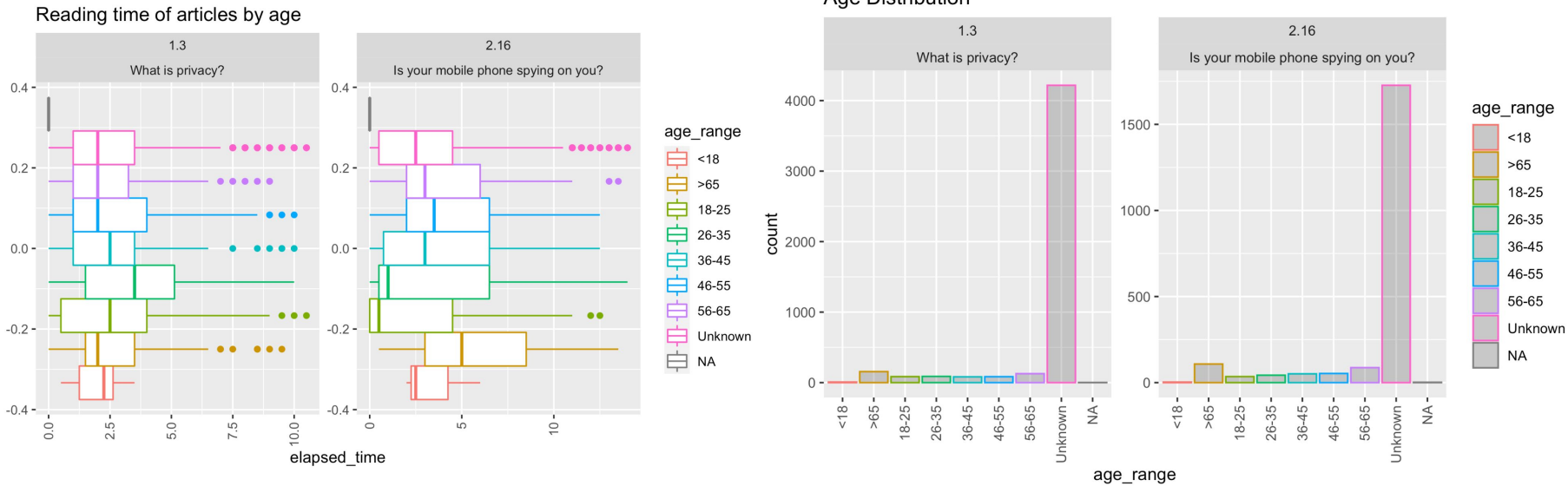
Reading Time



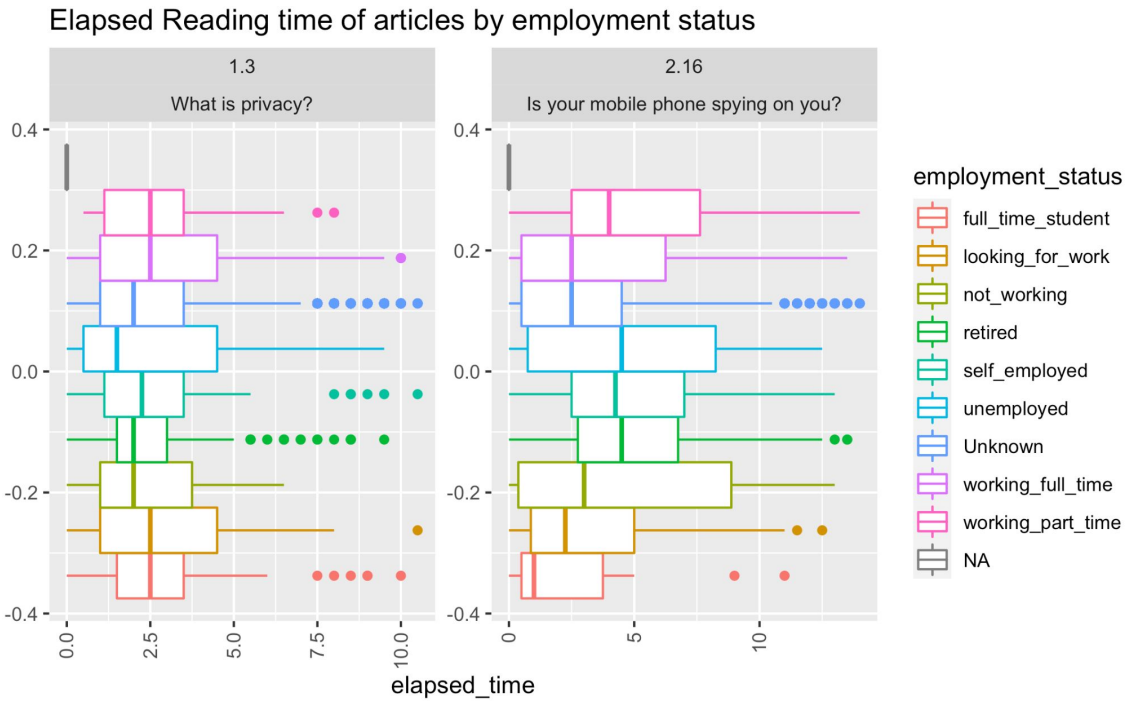
Average Article Reading Time



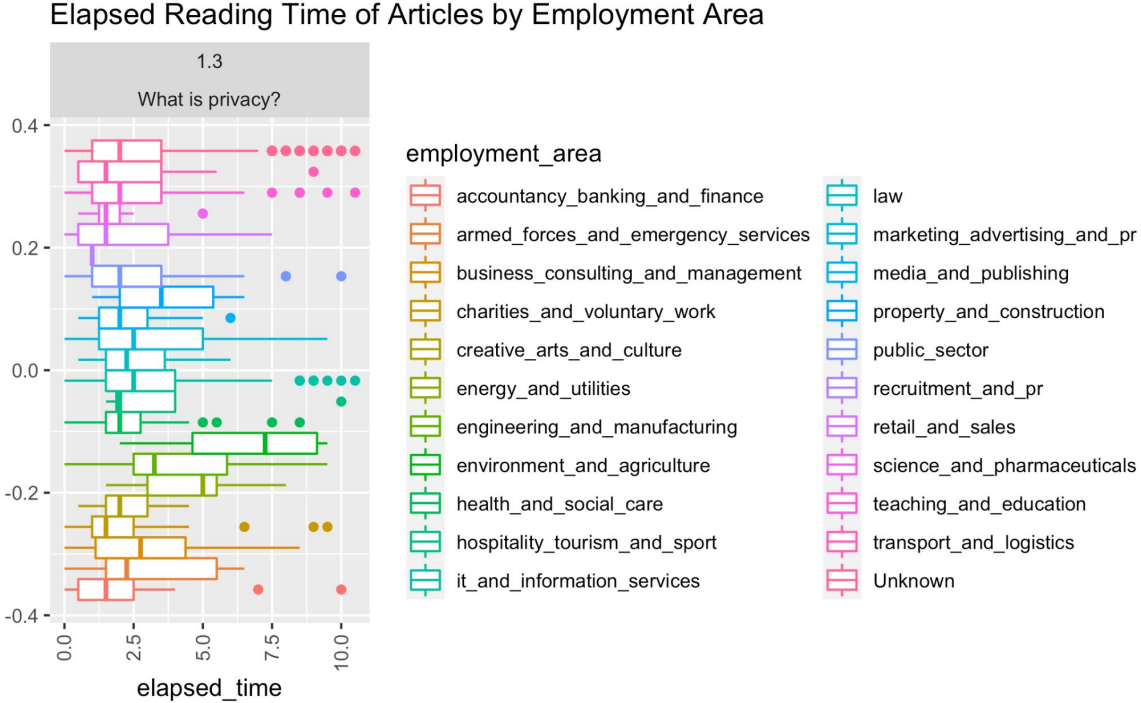
Age


Reading time of articles by age
1.3
What is privacy?
2.16
Is your mobile phone spying on you?
0.4
0.2
0.0
-0.2
-0.4
0.0
2.5
5.0
7.5
10.0
0
5
10
elapsed_time
age_range
<18
>65
18-25
26-35
36-45
46-55
56-65
Unknown
NA
Age Distribution
1.3
What is privacy?
2.16
Is your mobile phone spying on you?
count
4000
3000
2000
1000
0
1500
1000
500
0
<18
>65
18-25
26-35
36-45
46-55
56-65
Unknown
NA
age_range
age_range
<18
>65
18-25
26-35
36-45
46-55
56-65
Unknown
NA

Employment Status

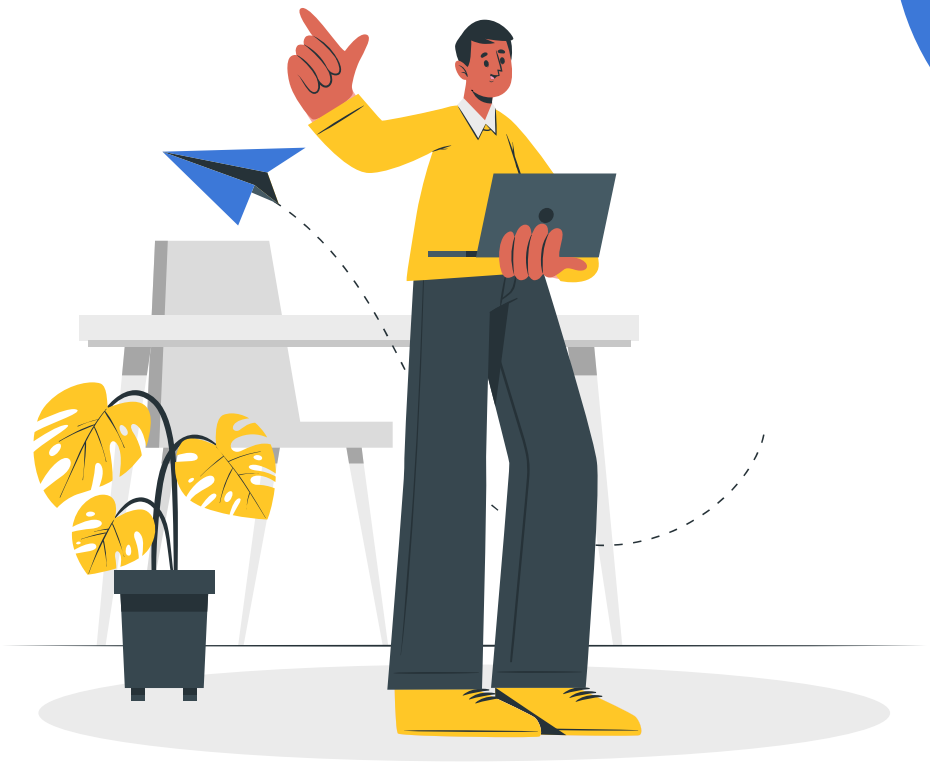


Employment Area

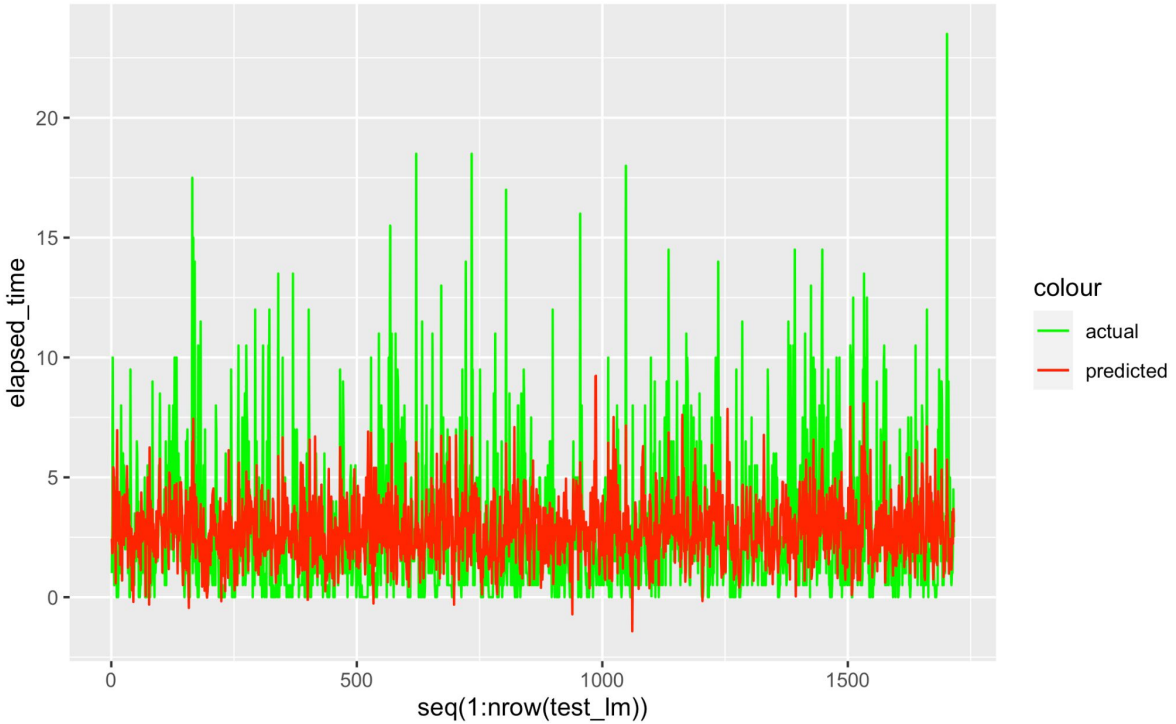


Model

Predict the estimated reading time



Linear Regression



RMSE: 2.60 min

Linear Regression