loqbox

University of
BRISTOL

BBC
TV

ANALYSING TEXT DATA ACROSS THE BUSINESS USING INNOVATIVE DATA SCIENCE TECHNIQUES

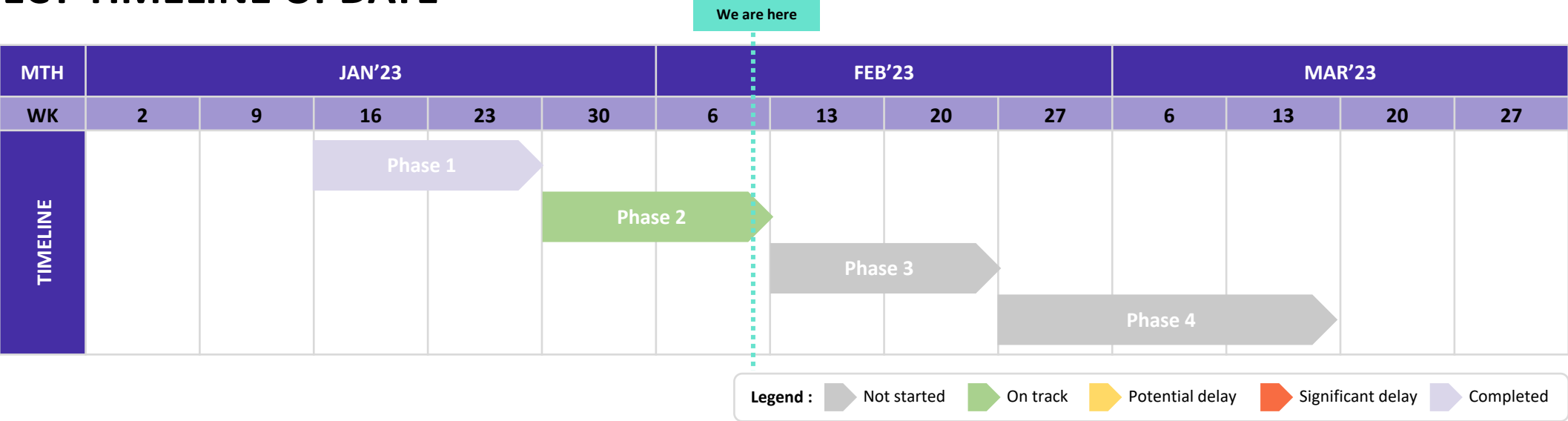## BI-WEEKLY STATUS UPDATE

09 FEBRUARY 2023

AGENDA

1. **PROJECT TIMELINE UPDATE**

2. **DATA CLEANING**

3. **DATA ANALYSIS METHODOLOGIES**

1

PROJECT TIMELINE UPDATE

We are here

MTH	JAN'23						FEB'23				MAR'23		
WK	2	9	16	23	30	6	13	20	27	6	13	20	27

Phase 1

Phase 2

Phase 3

Phase 4

Legend : Not started | On track | Potential delay | Significant delay | Completed

Phase	Activities & Deliverables
Phase 1 Initial data insights	• Collect the data. • Initial review of data set – identify insights and concerns to discuss with client. • Initiate data cleaning phase by creating a collaborative Python workbook.
Phase 2 Data processing	• Finalise data cleaning phase and initiate data analysis phase. • Identify key methodology for analysis and visualisation.
Phase 3 Data visualisation	• Finalise data analysis phase. • Integrate predictive and prescriptive analytic techniques to model an optimal membership scheme. • Finalise modelling analytics phase.
Phase 4 Project report and presentation	• Prepare and finalise project report and presentation. • Offer business insights and recommendations to client.

3

AGENDA

1. **PROJECT TIMELINE UPDATE**

2. **DATA CLEANING**

3. **DATA ANALYSIS METHODOLOGIES**

2

DATA CLEANING

DETAILS

1 OUTLIER REMOVAL

- Observations of users who did not activate any products were removed.
  - Deleted **371** observations.

activated_save	activated_spend	activated_rent	activated_coach
0	0	0	0

2 DATA CONSISTENCY

- Observations with multiple reasons in the "update_reason" column were split into rows.
  - Split the original data into a total of **55,433** observations.

customer_id	update_reason
7027	[{"value": "0", "reason": "My current plan is too expensive"}, {"value": "2", "reason": "I'm not taking advantage of all the benefits of my current plan"}]

➡

customer_id	update_reason
7027	{"value": "0", "reason": "My current plan is too expensive"}
7027	{"value": "2", "reason": "I'm not taking advantage of all the benefits of my current plan"}

- Three separate columns for **"value"**, **"reason"** and **"note"** were created for "update_reason".

3 DATA CLASSIFICATION

- The data was divided into categories based on the following criteria:
  - Downgraded date time – before or after **20th July 2022 at 14:30** (the day that a downgrading screen was changed).
  - The information in the "update_reason" column – whether it was written by the user or not (i.e. **note and value**).
  - The interval time before the users downgraded – **14 days**.

INITIAL INSIGHTS

- **581** customers had not activated loqbox save before downgrading – what this tells us?
- **7993** customers gave no reason.

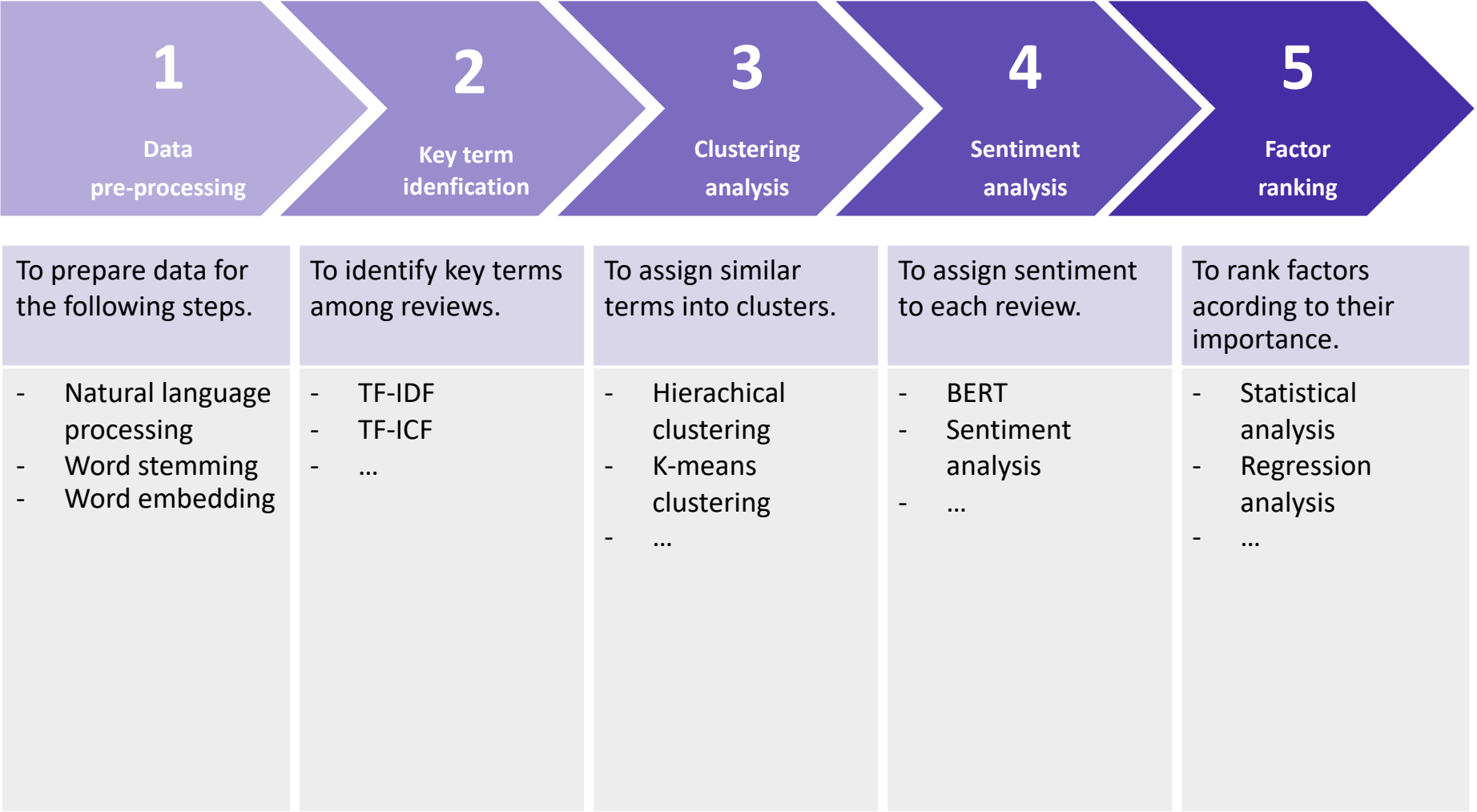Remark: The initial data set contains 45,064 observations from 6th July 2021 to 16th December 2022.

AGENDA

3

1. **PROJECT TIMELINE UPDATE**

2. **DATA CLEANING**

3. **DATA ANALYSIS METHODOLOGIES**

DATA ANALYSIS METHODOLOGIES

CUSTOMER REVIEW ANALYSIS FRAMEWORK

	1 Data pre-processing	2 Key term idenfication	3 Clustering analysis	4 Sentiment analysis	5 Factor ranking
OBJECTIVE	To prepare data for the following steps.	To identify key terms among reviews.	To assign similar terms into clusters.	To assign sentiment to each review.	To rank factors acording to their importance.
ANALYTICAL TECHNIQUES	- Natural language processing - Word stemming - Word embedding	- TF-IDF - TF-ICF - …	- Hierachical clustering - K-means clustering - …	- BERT - Sentiment analysis - …	- Statistical analysis - Regression analysis - …

RECOMMENDATIONS

APPENDIX I – SAMPLE DATAFRAME 1

	customer_id	upgraded_at	downgraded_at	activated_save	activated_rent	activated_coach	update_reason	value	reason	note
0	18	2022-06-06 11:05:27.000	2022-11-16 20:34:41.000	1	1	0	[{"value": "1", "reason": null}]	1	None	NaN
1	252	2022-02-12 09:23:22.000	2022-03-14 16:17:01.000	1	0	0	[{"note": "`Member wants to proceed with Save ...	2	NaN	`Member wants to proceed with Save only
2	267	2022-03-01 08:49:00.000	2022-05-08 12:37:56.000	1	0	0	[{"value": "0", "reason": "My current plan is ...	2	I'm not taking advantage of all the benefits o...	NaN
2	267	2022-03-01 08:49:00.000	2022-05-08 12:37:56.000	1	0	0	[{"value": "0", "reason": "My current plan is ...	1	My current plan isn't good value for money	NaN
2	267	2022-03-01 08:49:00.000	2022-05-08 12:37:56.000	1	0	0	[{"value": "0", "reason": "My current plan is ...	0	My current plan is too expensive	NaN
...	...	...	...	...	...	...	...	...	...	...
45059	784773	2022-12-16 11:28:09.000	2022-12-16 11:43:52.000	1	0	0	[{"value": "1", "reason": null}, {"value": "2"...	2	None	NaN
45060	784819	2022-12-16 12:11:55.000	2022-12-16 12:14:22.000	1	0	0	[{"value": "5", "reason": null}]	5	None	NaN

APPENDIX II – SAMPLE DATAFRAME 2

	customer_id	upgraded_at	downgraded_at	activated_save	activated_spend	activated_rent	activated_coach	update_reason	values	0	1	2	3	4	5	6	7
0	18	2022-06-06 11:05:27.000	2022-11-16 20:34:41.000	1	0	1	0	[{"value": "1", "reason": null}]	1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
1	252	2022-02-12 09:23:22.000	2022-03-14 16:17:01.000	1	0	0	0	[{"note": "`Member wants to proceed with Save ...	2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
2	267	2022-03-01 08:49:00.000	2022-05-08 12:37:56.000	1	0	0	0	[{"value": "0", "reason": "My current plan is ...	0,1,2	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
3	332	2022-08-05 23:25:15.000	2022-08-25 15:52:59.000	1	1	0	0	[{"value": "1", "reason": null}, {"value": "2"...	1,2,4	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0
4	445	2022-05-27 07:41:27.000	2022-05-27 07:43:36.000	1	0	0	1	[{"value": "0", "reason": "My current plan is ...	0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
44688	784693	2022-12-16 09:52:04.000	2022-12-16 10:34:04.000	1	0	0	0	[{"value": "1", "reason": "It's sneaky to offe...	1,2,3,4,5	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
44689	784773	2022-12-16 11:28:09.000	2022-12-16 11:43:52.000	1	0	0	0	[{"value": "1", "reason": null}, {"value": "2"...	1,2	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0
44690	784819	2022-12-16 12:11:55.000	2022-12-16 12:14:22.000	1	0	0	0	[{"value": "5", "reason": null}]	5	0.0	0.00	0.0	0.0	0.0	1.0	0.0	0.0
44691	784930	2022-12-16 14:14:46.000	2022-12-16 14:23:11.000	1	1	0	1	[{"value": "2", "reason": null}]	2	0.0	0.01	1.0	0.0	0.0	0.0	0.0	0.0
44692	784991	2022-12-16 15:24:35.000	2022-12-16 15:31:53.000	1	0	0	0	[{"value": "5", "reason": null}]	5	0.0	0.0	0.0	0.0	0.01	1.0	0.0	0.0

Q&A

THANK YOU