Department of Electronic and Electrical Engineering

EE1622 Data and Information
Boot camp Brief (SPSS section)

Table of Contents

Title	Twitter analytics
Module Leader	Prof Tatiana Kalganova
Prerequisites	All Mandatory Lab Tutorials
Submission Deadline	N/A
Feedback by	During Boot Camp
Contribution to overall module assessment	N/A
Indicative student time working on assessment	16 Hours
Word or Page Limit (if applicable)	n/a (enter answers as required into the provided Submission form)

Main Objective of the Boot camp
In this assessment, you will demonstrate module learnings outcomes one (LO1), two (LO2) and three (LO3) within the context of a Twitter data collection. More specifically you will download a set of Tweets data and prepare this data for further analysis using SPSS (e.g. adding metadata, filtering unwanted cases, deriving new variables and exporting the data table to a Database format). You will then perform a series of statistical analyses on this data to answer various questions.

This assessment prepares you for the Class Test and it is mandatory to attend each of the three boot camp days. Failing to attend will exclude you to sit the class test. It is therefore critical that you attempt this coursework to the best of your ability.

Description of the Assessment
The assessment consists of the following tasks:

1. ***Completion of the mandatory Lab Tutorials:*** These were released during Term 1 and can be found on Blackboard in the Laboratory section. Completion of a tutorial can be demonstrated by taking the associated quiz on Blackboard.
2. ***Data table preparation:*** You will download a set of Twitter data (Tweets), from Blackboard. You will find the dataset in the boot camp folder as Brexit _Tweets1.xlsx. These tweets cover the 8 days up to and including the EU Referendum on 23rd June 2017. You will then import these into an SPSS data table and derive a number of database tables from this data. The **Data Checklist** is described in the

EE1622 Data and Information
Submission Form

Submission Form below which specifies the criteria that your data files must meet.

Description of the Task
The aim of this task is to prepare and analyse a large set of tweets collected in the 8 days leading up to and including the EU Referendum on 23rd June 2016. The query used was "EU AND (brexit OR remain OR leave or

stay)". The raw dataset is a text file (tab delimited) containing a header row plus just over 90,000 rows where each row is a distinct tweet.

Over the following pages, you be presented with a number of processing and analysis tasks. Processing tasks include filtering rows and creating new variables, using both SPSS and SQL Lab tool. Analysis tasks will require you to use either SPSS to compute statistics, or SQL Lab tool to select results tables.

You can download this dataset from Blackboard in the same folder you downloaded this document from.

Import Task

Load the data into SPSS using the Text Import Wizard (File → Open → Data) making sure of the following:

- Data is delimited by tab
- Variable names are extracted from the top row
- The number of variables is correct (i.e. no fields are split across multiple variables)
- All variables are named correctly based on the data dictionary shown below. Note that many names in the header vary slightly from the data dictionary definitions shown below and will need to be changed either during import or immediately after)
- All variables are of the correct type, following the data dictionary. In particular, ensure that the DateTime values are preserved intact and the datatype is Date/Time (not String)

Data Dictionary

Variable name	Description	Type/measure	Coding notes
dateTime	Date and Time of tweet	DateTime/Ordinal	dd-mmm-yyyy hh:mm:ss [using Tweetcatcher to convert into SPSS compatible format before creating the SPSS table]
screenName	Screen name of user	String/Nominal	
tweet	Content of tweet	String/Nominal	
tweetID	Unique identifier of tweet	Numeric/Ordinal	
realName	Real name of user	String/Nominal	
followers	Number following user	Numeric/Scale	
friends	Number followed by user	Numeric/Scale	
location	Number of lists subscribed to	String/Nominal	
utcOffset	Time difference relative to UTC/GMT	Numeric/Scale	There is a finite set of timezone strings
geoCoords	Geographical coordinates provided at time tweet sent	String/Nominal	Two sub-strings (longitude, latitude) delimited by a space)
userTweets	Number of tweets posted since account opened	Numeric/Scale	
reTweets	Number of retweets of the original tweet	Numeric/Scale	
linkURL	URL of link embedded within the tweet	String/Nominal	
sentPos	Positive sentiment rating	Numeric/Ordinal	Range of 1 to 5
sentNeg	Negative sentiment rating	Numeric/Ordinal	Range of -1 to -5
userBio	User profile description (biography)	String/Nominal	

Any derived variables resulting from the questions should also be correctly named/typed and included in your submitted SPSS data table and/or the database tables.

THRESHOLD CHECK LIST

The data check list is provided here to help you to make sure you've included all of the required data files in the correct format. A member of the teaching team will check your files and use the checklist to determine whether you have met the minimum requirements for a pass.

Your Name:	**Brunel Id:**	

LO1: SPSS	**Marker comment**	**Outcome**
Data table has all 16 variables present with correct name, type and measure attributes as specified in the data dictionary.		**Y/N**
Table contains data and no columns are empty		**Y/N**
Table contains the same number of rows as declared in the students answer to LO1 Q? (following removal of non-UK tweets, see below)		Y/N
Data table includes the derived variable ffRatio (see below)		Y/N