Assignment 3

Practical Introduction to Data Science (SS1-SEM2)

Candidate Exam Number: B136325

Department of Informatics
University of Edinburgh
Scotland
9th June 2019

Contents

Question 1

Introduction

The demonstration will be structured with regard to the CRISP-DM [1] data science project methodology, which describes such projects in terms of the following six sequential phases: Business Understanding, Data Understanding, Data Preparation, Analysis, Evaluation and Deployment. In addition, the CRISP-DM Data Understanding phase will contain the data cleaning components of the Statistical Value Chain, namely, Technically Correct and Consistently Correct data. Nevertheless, the first of the CRISP-DM phases, Business Understanding, will be presented immediately below.

Business Understanding

The goal of this question will be to demonstrate that weather data can be clustered into two or more groups with each containing 'similar' weather conditions. The data contains results from thirty six monitoring stations across the UK (as made available online by the Met Office). While the process of retrieving and then cleaning the data will be described within the Data Understanding stage further below, it is worth noting that all of the process described within this answer have been implemented as an R package called **pids.wellbeing.weather**. The package's name begins with an abbreviation of the Practical Introduction to Data Science course, namely **pids**, followed by the names of the two primary datasets within the package: **wellbeing.weather**. Further information about the package will be provided immediately below.

pids.wellbeing.weather

GitHub

The **pids-wellbeing-weather** package has been made available on GitHub [2], using this student's exam number, **b136325**, as the associated GitHub account name. In addition, the package makes use of Packrat [3] dependency management. This means that the package contains all of its dependencies. This approach increases the package's size. Importantly, however, it ensures portability, and the package can be cloned (and used immediately) from GitHub with the following command.

```
git clone https://github.com/b136325/pids.wellbeing.weather.git
```

RStudio

Alternatively, the package can be installed into RStudio using the commands below.

```
install.packages("devtools")
library(devtools)
install_github("b136325/pids.wellbeing.weather")
```

Structure

The **pids-wellbeing-weather** package has been structured in accordance with R best practice [4]. For example, raw data can be found within the **./data-raw/** directory. In addition, both *technically* and *consistently correct* data can be found within the **./data/** directory. The code, which was linted using Lintr [5], and which was semantically versioned [6] using Git [7], can be found within the **./R/** directory.

Files and exported function names

Broadly, each code file exports only one function, and the name of the exported function accords with the associated file name. For example, the first code file relating to question 1, which downloads raw data, and whose address can be found directly below, exports a function called *question_1_001_svc_raw-data*. In almost all cases, the names of exported functions match their associated files. This approach was adopted to ensure that exported functions could be found quickly within the code.

```
./R/question-1-001-svc-raw-data.R
```

Exported function names and questions

In addition, the name of each export function begins with the related question number. For example, the exported function ***question_1***_001_svc_raw-data, as described above, relates to question 1. In contrast, the exported function ***question_2***_007_analysis_knn, which provides K Nearest Neighbours analysis, relates to question 2. This approach has been adopted for all exported functions, and it provides a structure for the code in relation to the questions.

Function sequence number

The name of each exported function (and the name of each associated code file) also contains a three digit number. This is the **function sequence number**. It can be found immediately after the question number, and it represents the function's sequence of use in relation to a specific question. For example, the exported function *question_1_**001**_svc_raw-data*, as, described above, is the first function relating to question 1. In contrast, *question_1_**036**_eva_charts_hier_latitude_longitude_min_temp* is the 36th exported function relation to the same question.

Function names and question phases

Lastly, each exported function name contains a **question phase** description. This can be found immeditaley after the **function sequence number**, and it describes the phase (or part) of a question to which the function is related. For example, the exported function *question_1_023_**analysis**_hierarchical*, which resturns a hierarchical cluster, is a part of the analysis phase for question 1. In contrast, *question_2_003_**eda**_charts_weather_features* is an exported function, which returns a chart of weather features, and which is related to the Exploratory Data Analysis (EDA) phase of question 2. A full list of the **question phases** used within the exported function names can be found in Appendix .

Summary

It is suggested that the function naming conventions describe above would be too brittle for use within an ongoing project. However, the adoption of such conventions should simplify references to functions (within this document) and their use in the **pids-wellbeing-weather** package. Within the remainde of this document, the code to call an associaetd exported function (or functions) will be listed at the end of each phase (or part) of a question. In each case, file paths will also be proved.

Data Understanding

The aim of this phase of question 1 is to understand the data. In order to begin such understanding, the raw data must first be retrieved and processed. These tasks will be undertaken with regard to the three data cleaning stages of the Statistical Value Chain, which will be descibed in further detail below.

Statistical value chain

The three data cleaning stages of the Statistical Value Chain provide a sequential structure for understanding the processes involved in transforming raw data into data ready to be analysed. The three stages are as follows: (1) raw data; (2) *technically complete* data; and (3) *consistently complete* data. In this case, the ouput from the first stage would be raw data retrieved from an extenal source. The raw data would then be used as the input to the second stage, *technically complete*. The ouput from the second stage would be data with consistent column (or feature) naming, along with the use of appropriate variable data types per column. From the perspective of the R language [8], examples of data types (that could be applied to such columns) include (but are not limited to) *character*, *date* and *double*. In addition, *technically complete* data should have addressed *null*, *empty* and similarly problematic values with the data. Such technically complete data would then be passed to the final stage, *consistently complete*, where the internal consistency (and structure) of the data would be addressed. The output from the last of those stages would be data ready for analysis. Having now outlined the three data cleaning stages of the Statistical Value Chain (as adopted within the Data Understanding phase of question 1), the first stage, that of raw data, will now be examined in greater detail.

Raw data: weather and well-being

This implementation of the raw data stage of the Statistical Value Chain involved downloading the required *weather* and *wellbeing* files into the **pids-wellbeing-weather** package, where the files were stored in the **./data-raw/weather** and **./data-raw/wellbeing** directories, respectively. The web addresses (from which the files were downloaded) were constructed dynamically using constants defined within the package. The defined constants can be found within the package at **./R/constants.R**. The function associated with this process (which performs the download) has been exported from the package, and it can be run using the command below. It returns a list of files that have been downloaded successfully, along with any that have failed, as illustrated in Appendix .

```
pids.wellbeing.weather::question_1_001_svc_raw_data()
```

The function described above can be found at the following address within the **pids.wellbeing.weather** package.

```
./R/question-1-001-svc-raw-data.R
```

Technically correct: weather .txt data

This implementation of the second Statistical Value Chain data cleaning stage, that of *technically correct*, has been divided into five parts. This approach ensures that each of those parts perform a relatively small change to the data, making testing easier than might otherwise have been the case. Nevewrtheless, in order to ensure a strong data related *separation of concerns*, such that, no amendments should be made to the data downloaded into the **./data-raw** directory, this part of the technically correct stage copies the raw weather data into the **./data** directory. More specifically, the weather data is copied into a directory representing the first stage of the technically correct process: **./data/weather/stage-010-technically-correct-text** with **.txt** file extension. In addition, a closing character is added to each file, which facilitates subsequent processing. The function associated with this first part of the *technically complete* stage has been exported from the package. It can be run using the command below, and it returns a list containing the destination paths of successfully moved files.

```
pids.wellbeing.weather::question_1_002_svc_tech_weather_txt()
```

The function described above can be found at the following address within the **pids.wellbeing.weather** package.

```
./R/question-1-002-svc-tech-weather-txt.R
```

Technically correct: weather .dsv data

The second part of *technically correct* stage involved copying and then transforming the **.txt** files (as described immediately above) into **.dsv**, white space delimited equivalents. In addition, non column related header items were removed, and all of the files were transformed into containg a standard number of columns: eight. The latter transformation enabled the files to be reliably converted into the aforementioned delimited format. Lastly, simple invalid data items, such as — were converted into a common R language data type, namely **NA**. The delimited files were saved to the following directory **./data/weather/stage-011-technically-correct-dsv**. The function associated with this part of the *technically correct* stage, and which performs the tasks described immediately above, has been exported from the **pids-wellbeing-weather** package. It can be run using the command below, and it returns a list containing the destination paths of the successfully copied and transformed files.

```
pids.wellbeing.weather::question_1_003_svc_tech_weather_dsv()
```

The function described above can be found at the following address within the **pids.wellbeing.weather** package.

```
./R/question-1-003-svc-tech-weather-dsv.R
```

Technically correct: weather data frames

The third part of the *technically correct* stage transformed the .dsv files into R language data frames. The data frames were then saved in **.Rds** format within the **data/weather/stage-012-technically-correct-dataframe** directory. The function associated with this process has been exported from the package and it can be run using the command below. It function returns a list containing the destination paths of the successfully saved files.

```
pids.wellbeing.weather::question_1_004_svc_tech_weather_df()
```

The function described above can be found at the following address within the **pids.wellbeing.weather** package.

```
./R/question-1-004-svc-tech-weather-df.R
```

Technically correct: weather data frames completet

The fourth part of the *technically correct* stage copied and transformed the **.Rds** files (described above) . It ensured that the data frames had lower case column names, and underscores replaced hyphens or spaces in such names. It also ensured that appropriate data types were applied to the columns (or features). The transformed data frames were saved within the **./data/weather/stage-013-technically-correct-complete** directory. The function associated with this process has been exported and it can be run using the command below. It function returns a list containing the destination paths of the successfully saved files.

```
pids.wellbeing.weather::question_1_005_svc_tech_weather_complete()
```

The function described above can be found at the following address within the **pids.wellbeing.weather** package.

```
./R/question-1-005-svc-tech-weather-complete.R
```

Technically correct: weather single data frame

The final part of the *technically correct* stage transformed the **.Rds** files (as described above) into a single data frame stored within **./data/weather/stage-014-technically-complete-single-dataframe**. The function associated with this process has been exported and it can be run using the command below. It returns the destination path of the successfully saved file.

```
pids.wellbeing.weather::question_1_006_svc_tech_weather_single_df()
```

The function described above can be found at the following address within the **pids.wellbeing.weather** package.

```
./R/question-1-006-svc-tech-weather-single-df.R
```

Consistently correct

Summary Type	Sunshine (hours)	Rain (mm)	Temp max (c)	Temp min (c)
Mean	119.119	75.032	12.816	5.998
Min	89.526	45.803	9.496	2.734
Max	155.310	151.117	14.968	8.386
SD	16.196	26.958	1.343	1.194

Table 1: Mean, min, max and SD for the weather features (3sf)

```
pids.wellbeing.weather::question_1_007_svc_cons_summary()
pids.wellbeing.weather::question_1_008_svc_cons_grouped_data()
pids.wellbeing.weather::question_1_009_svc_cons_grouped_data_summary()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-1-007-svc-cons-summary.R
./R/question-1-008-svc-cons-grouped-data.R
./R/question-1-009-svc-cons-grouped-data-summary.R
```

Exploratory data analysis (EDA)

Box and whisker

```
pids.wellbeing.weather::question_1_010_eda_charts_box_whisker_hours_sun()
pids.wellbeing.weather::question_1_010_eda_charts_box_whisker_hours_rain()
pids.wellbeing.weather::question_1_010_eda_charts_box_whisker_max_temp()
pids.wellbeing.weather::question_1_010_eda_charts_box_whisker_min_temp()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question_1_010_eda_charts_box_whisker_hours_sun.R
./R/question_1_010_eda_charts_box_whisker_hours_rain.R
./R/question_1_010_eda_charts_box_whisker_max_temp.R
./R/question_1_010_eda_charts_box_whisker_min_temp.R
```

Remove outliers

Summary Type	Sunshine (hours)	Rain (mm)	Temp max (c)	Temp min (c)
Mean	118.43	69.602	12.816	5.998
Min	91.320	47.399	9.496	2.734
Max	149.349	127.368	14.968	8.386
SD	14.032	18.998	1.343	1.194

Table 2: Mean, min, max and SD for the weather features (3sf)

```
pids.wellbeing.weather::question_1_011_eda_remove_outliers()
pids.wellbeing.weather::question_1_012_eda_remove_outliers_summary()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-1-011-eda-remove-outliers.R
./R/question-1-012-eda-remove-outliers-summary.R
```

Weather variables by latitude

```
pids.wellbeing.weather::question_1_013_eda_charts_latitude()
```

Within the *pids.wellbeing.weather* package the function described above can be found at the following address:

```
./R/question-1-013-eda-charts-latitude.R
```

Weather variables by latitude and longitude

```
pids.wellbeing.weather::question_1_014_eda_charts_longitude_latitude_hours_sun()
pids.wellbeing.weather::question_1_015_eda_charts_longitude_latitude_rain()
pids.wellbeing.weather::question_1_016_eda_charts_longitude_latitude_max_temp()
pids.wellbeing.weather::question_1_017_eda_charts_longitude_latitude_min_temp()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-1-014-eda-charts-longitude-latitude-hours-sun.R
./R/question-1-015-eda-charts-longitude-latitude-rain.R
./R/question-1-016-eda-charts-longitude-latitude-max-temp.R
./R/question-1-017-eda-charts-longitude-latitude-min-temp.R
```

Cluster tendency

The clustering tendency of the data has been calculated using the Hopkins statistic (H). It assesses the probability that the data contains non random structures. The statistic has been calculated using the *factoextra* dependency. Using the data with outliers removed, the result of H was **0.352**.

(1) high score–¿uniform distribution–¿no cluster

(2)low score–¿not uniform distribution-/-¿(may be not)cluster.

The function used to generate this statistic can be run using the command below:

```
pids.wellbeing.weather::question_1_018_prep_cluster_tendency()
pids.wellbeing.weather::question_1_018_prep_cluster_tendency(show_chart = TRUE)
```

Within the *pids.wellbeing.weather* package the function described above can be found at the following address:

```
./R/question-1-018-prep-cluster-tendency.R
```

VAT

```
pids.wellbeing.weather::question_1_019_prep_charts_vat()
```

Within the *pids.wellbeing.weather* package the function described above can be found at the following address:

```
./R/question-1-019-prep-charts-vat.R
```

Data preparation

K choice

```
pids.wellbeing.weather::question_1_020_prep_k_choice()
```

Within the *pids.wellbeing.weather* package the function described above can be found at the following address:

```
./R/question-1-020-prep-k-choice.R
```

Algorithm choice

Model type	Speed	Independence	Unequal cluster sizes	Unusual density
K-Means			No	No
K-Mediods			Yes	No
Hierarchical				
Distribution based			Yes	Yes

Table 3: Summary of clustering algorithms by feature)

```
pids.wellbeing.weather::question_1_021_prep_algorithm_choice()
```

Within the *pids.wellbeing.weather* package the function described above can be found at the following address:

```
./R/question-1-021-prep-algorithm-choice.R
```

Linkage choice

```
pids.wellbeing.weather::question_1_022_linkage_choice()
```

Within the *pids.wellbeing.weather* package the function described above can be found at the following address:

```
./R/question-1-022-linkage-choice.R
```

Analysis

Hierarchical

```
pids.wellbeing.weather::question_1_023_analysis_hierarchical()
pids.wellbeing.weather::question_1_024_analysis_charts_dendrogram()
pids.wellbeing.weather::question_1_025_analysis_attach_pruned_cluster_values()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-1-023-analysis-hierarchical.R
./R/question-1-024-analysis-charts-dendrogram.R
./R/question-1-025-analysis-attach-pruned-cluster-values.R
```

k-means

```
pids.wellbeing.weather::question_1_026_analysis_kmeans()
pids.wellbeing.weather::question_1_027_analysis_charts_sum_squares()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-1-026-analysis-kmeans.R
./R/question-1-027-analysis-charts-sum-squares.R
```

Evaluation

Hierarchical

```
pids.wellbeing.weather::question_1_032_eval_charts_hierarchical_latitude()
pids.wellbeing.weather::question_1_033_eval_charts_hier_latitude_longitude_hours_
pids.wellbeing.weather::question_1_034_eval_charts_hier_latitude_longitude_rain_m
pids.wellbeing.weather::question_1_035_eval_charts_hier_latitude_longitude_max_te
pids.wellbeing.weather::question_1_036_eval_charts_hier_latitude_longitude_min_te
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-1-032-eval-charts-hier-latitude.R
./R/question-1-033-eval-charts-hier-latitude-longitude-hours-sun.R
./R/question-1-034-eval-charts-hier-latitude-longitude-rain-mm.R
./R/question-1-035-eval-charts-hier-latitude-longitude-max-temp.R
./R/question-1-036-eval-charts-hier-latitude-longitude-min-temp.R
```

Deployment

Question 2

Business Understanding

Data Understanding

Append latitude categories

```
pids.wellbeing.weather::question_2_001_bu_append_latitude_categories()
```

Within the *pids.wellbeing.weather* package the functions described above can be found at the following address:

```
./R/question-2-001-bu-append-latitude-categories()
```

Exploratory data analysis (EDA)

Number of weather stations per latitude category

```
pids.wellbeing.weather::question_2_002_eda_latitude_category_summary()$nrow
```

The function described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-002-eda-latitude-category-summary.R
```

Weather stations per latitude category

```
pids.wellbeing.weather::question_2_002_eda_latitude_category_summary()$weather_st
```

The function described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-002-eda-latitude-category-summary.R
```

Mean weather features per latitude category

```
pids.wellbeing.weather::question_2_002_eda_latitude_category_summary()$means
pids.wellbeing.weather::question_2_003_eda_charts_weather_features()
```

The function described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-002-eda-latitude-category-summary.R
./R/question-2-003-eda-charts-weather-features.R
```

Latitude category pairwise comparisons

```
pids.wellbeing.weather::question_2_004_eda_charts_latitude_category_pairwise()
```

The function described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-004-eda-charts-latitude_category-pairwise.R
```

Latitude and longitude per latitude category

```
pids.wellbeing.weather::question_2_005_eda_charts_latitude_longitude_hours_sun()
pids.wellbeing.weather::question_2_005_eda_charts_latitude_longitude_rain()
pids.wellbeing.weather::question_2_005_eda_charts_latitude_longitude_temp_max
pids.wellbeing.weather::question_2_005_eda_charts_latitude_longitude_temp_min
```

The functions described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-005-eda-charts-latitude-longitude.R
```

Data Preparation

Training and test data split

```
pids.wellbeing.weather::question_2_006_prep_split()
```

The function described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-006-prep-split.R
```

Model type	Speed	Normal	Independence	Conditional prob
K-NN Multiple Logistic Regression Naive Bayes				

Table 4: Summary of clustering algorithms by feature)

Algorithm choice

Analysis

K Nearest Neighbours

```
pids.wellbeing.weather::question_2_007_analysis_knn()
```

The function described above can be found within the *pids.wellbeing.weather* package at the following address:

```
./R/question-2-007-analysis-knn.R
```

Question 3

Business Understanding

Data Understanding

Statistical value chain

Technically correct: well-being

```
pids.wellbeing.weather::question_3_001_svc_tech_wellbeing()
```

```
./R/question-3-001-svc-tech-wellbeing.R
```

Consistently correct: weather and well-being

```
pids.wellbeing.weather::question_3_002_svc_cons_weather_add_boundaries()
pids.wellbeing.weather::question_3_003_svc_cons_weather_wellbeing_join()
pids.wellbeing.weather::question_3_004_svc_cons_weather_wellbeing_summary()
```

```
./R/question-3-002-svc-cons-weather-add-boundaries.R
./R/question-3-003-svc-cons-weather-wellbeing-join.R
./R/question-3-004-svc-cons-weather-wellbeing-summary.R
```

Exploratory data analysis (EDA)

```
pids.wellbeing.weather::question_3_005_eda_charts_wellbeing_by_region()
pids.wellbeing.weather::question_3_006_eda_charts_correlation()
```

```
./R/question-3-005-eda-charts-wellbeing-by-region.R
./R/question-3-006-eda-charts-correlation.R
```

Data Preparation

Analysis

```
pids.wellbeing.weather::question_3_007_analysis_regression()
pids.wellbeing.weather::question_3_008_analysis_regression_test()
```

```
./R/question-3-007-analysis-regression.R
./R/question-3-008-analysis-regression-test.R
```

Evaluation

```
pids.wellbeing.weather::question_3_009_eval_regression_summary()
pids.wellbeing.weather::question_3_010_eval_charts_predicted_happinesst()
```

```
./R/question-3-009-eval-regression-summary.R
./R/question-3-010-eval-charts-predicted-happiness.R
```

Bibliography

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0." `https://www.the-modeling-agency.com/crisp-dm.pdf`, January 2000. [Online].

[2] "pids.wellbeing.weather." `https://github.com/b136325/pids.wellbeing.weather/`. [Online].

[3] "Packrat." `https://rstudio.github.io/packrat/`. [Online].

[4] H. Wickham, "R pckages." `http://r-pkgs.had.co.nz/intro.html`. [Online].

[5] "Lintr." `https://github.com/jimhester/lintr`. [Online].

[6] "Semvar." `https://semver.org/`. [Online].

[7] "Git." `https://git-scm.com/`. [Online].

[8] "R data types." `https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures/`. [Online].

Appendices

A. Function names and question phases

QUESTION PHASES HERE

B. Raw data function output

IMAGE OF RAW DATA OUTPUT HERE