MATH2349 Data Preprocessing Assignment 2

*(* Last Updated 6.04.2018 )

**Weight:**20%

**Due date:**29 April, 2018, 23:59 AEST.

**Length:**Maximum 20 pages

**Feedback mode:** Feedback will be provided using Turnitin's inline marking tool and general comments.

This assignment requires you to apply different data manipulation techniques (like recode, filter, select, split, aggregate, and reshape the data) and justify data by detecting and handling missing values, outliers. This assignment is worth 20% and is due **29/04/2018**.

**Groups**

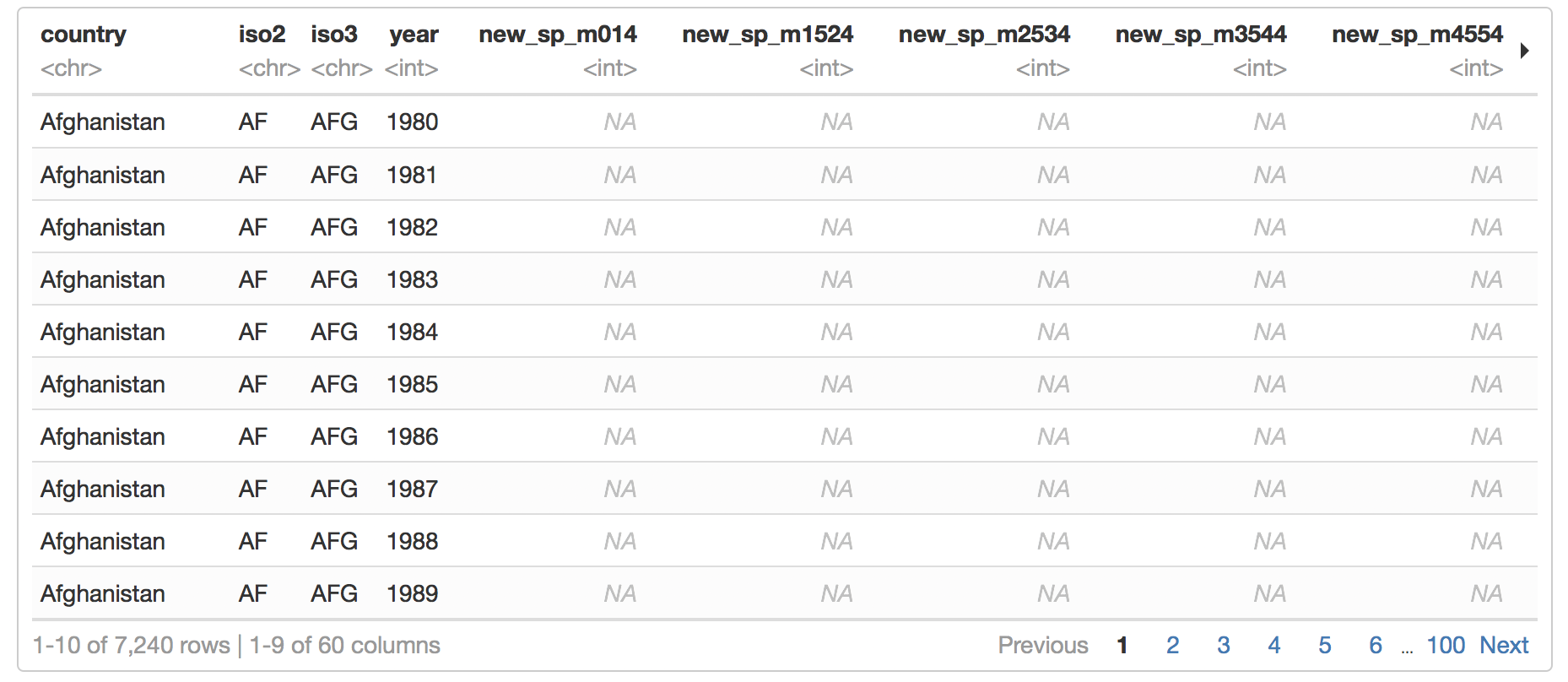
Students are permitted to work individually or in groups of up to 3 people for Assignment 2. **Each group must fill out the following form before 22/04/2018 to register their group details. After the deadline, group registrations won't be accepted.** Submit the details of your group here:

## [Individual and Group Registration Form](https://goo.gl/forms/Y89jLfWVbLUBbmjm2)

You will undertake ten different data manipulation and justification tasks using three different data sets. Here are the descriptions of the data sets used in this assignment:

WHO Data Description

Tasks 1 - 5 should be completed using the WHO data set ([available here](https://drive.google.com/drive/folders/15egkpiwwFPysQwoYIoY0ietAvLq61jeG?usp=sharing)). This data set contains tuberculosis (TB) cases reported between 1995 and 2013 sorted by country, age, and gender. The data comes in the 2014 World Health Organization Global Tuberculosis Report (source [www.who.int/tb/country/data/download/en/ )](http://www.who.int/tb/country/data/download/en/) and provides a wealth of epidemiological information. However, it would be difficult to work with the data as it is, as the data is not in a tidy format. Here is a portion of the data set:



**Variables in the WHO data**

WHO data set has a unique coding system for the variables. Columns five through sixty encode four separate pieces of information in their column names:

**1.** The first three letters of each column denote whether the column contains new or old cases of TB. In this data set, each column contains “**new**” cases.

**2.** The next two letters describe the type of case being counted. We will treat each of these as a separate variable.

* “**rel**” stands for cases of relapse
* “**ep**” stands for cases of extra-pulmonary TB
* “**sn**” stands for cases of pulmonary TB that could not be diagnosed by a pulmonary smear (smear negative)
* “**sp**” stands for cases of pulmonary TB that could be diagnosed be a pulmonary smear (smear positive)

**3.** The sixth letter describes the sex of TB patients. The data set groups cases by males (“**m**”) and females (“ **f** ”).

**4.** The remaining numbers describe the age group of TB patients. The data set groups cases into seven age groups:

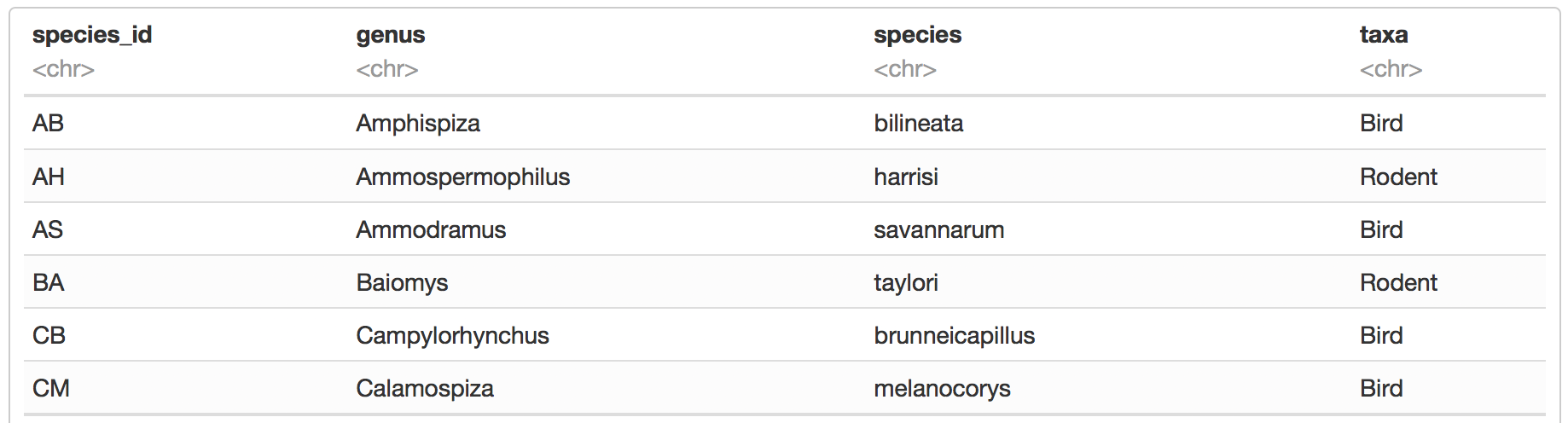
* “**014**” stands for patients that are 0 to 14 years old
* “**1524**” stands for patients that are 15 to 24 years old
* “**2534**” stands for patients that are 25 to 34 years old
* “**3544**” stands for patients that are 35 to 44 years old
* “**4554**” stands for patients that are 45 to 54 years old
* “**5564**” stands for patients that are 55 to 64 years old
* “**65**” stands for patients that are 65 years old or older

Note that the WHO data set is untidy, the data appears to contain values in its column names.

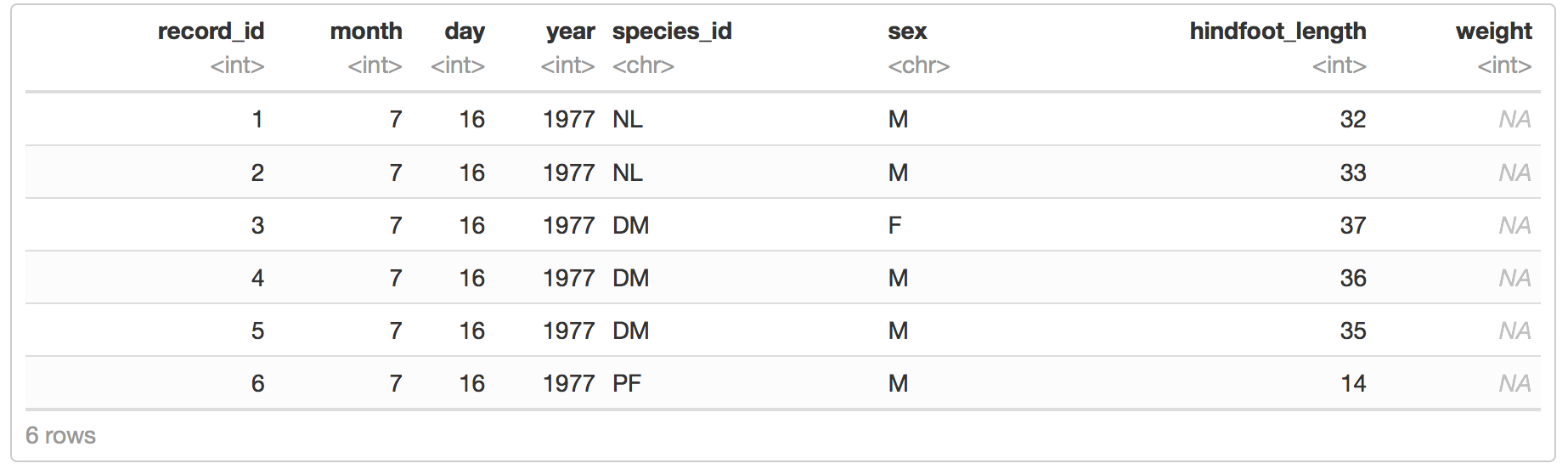
Species and Surveys Data Description

Tasks 6 - 10 will be completed using two relational sets that derived from a long-term study of animal populations in the Chihuahuan Desert (Data Source: http://dx.doi.org/10.6084/m9.figshare.1314459). The first data set, "**species.csv**", contains the information on the observed species where else the "**surveys.csv**" data lists all information related to the species available in the Chihuanuan Desert. Both data sets ([available here](https://drive.google.com/drive/folders/15egkpiwwFPysQwoYIoY0ietAvLq61jeG?usp=sharing)) have a common variable called “**species\_id**” connecting each other. The variables for Species and Surveys data sets are self-explanatory.

Species data (first six observations):



Surveys data (first six observations):

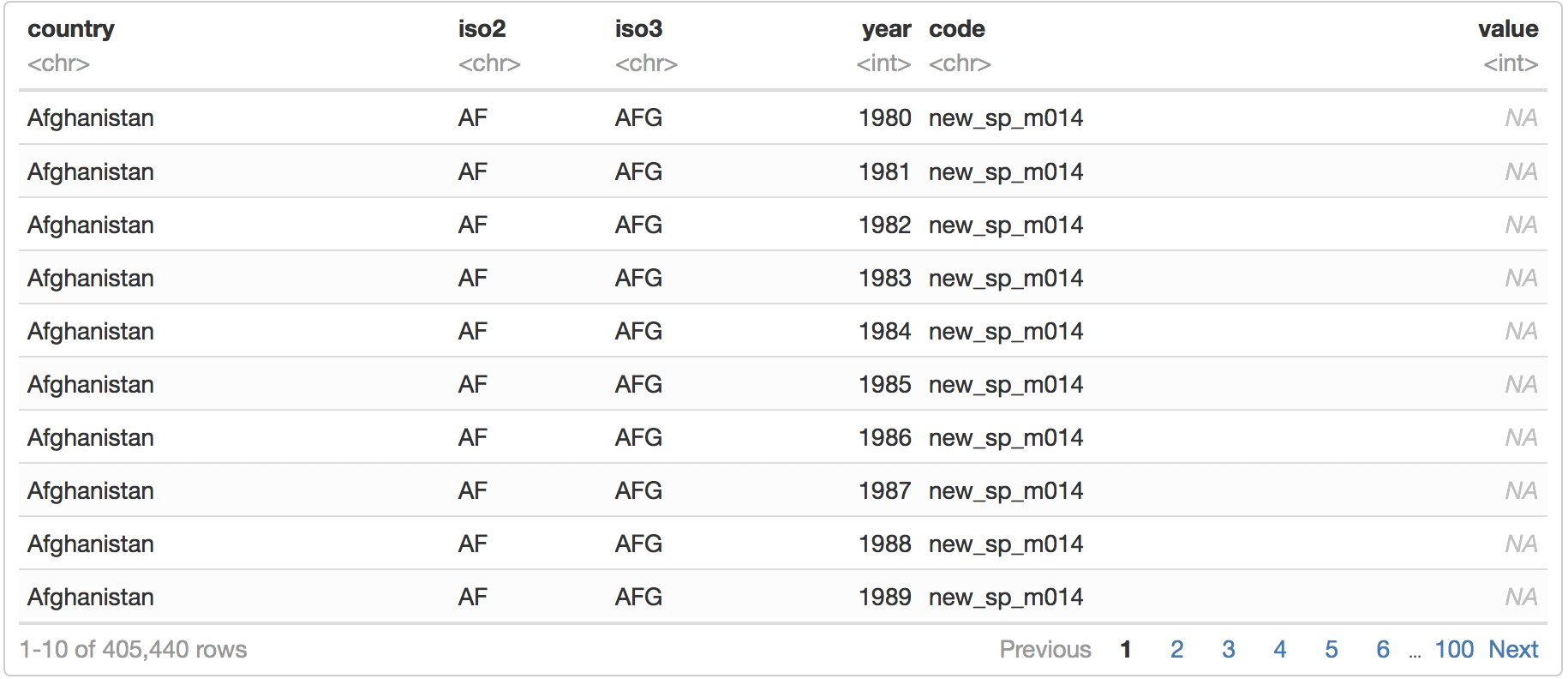


Assignment Tasks:

You will use WHO data set for Tasks 1- 5. Read the WHO data using an appropriate function and complete the tasks 1-5.

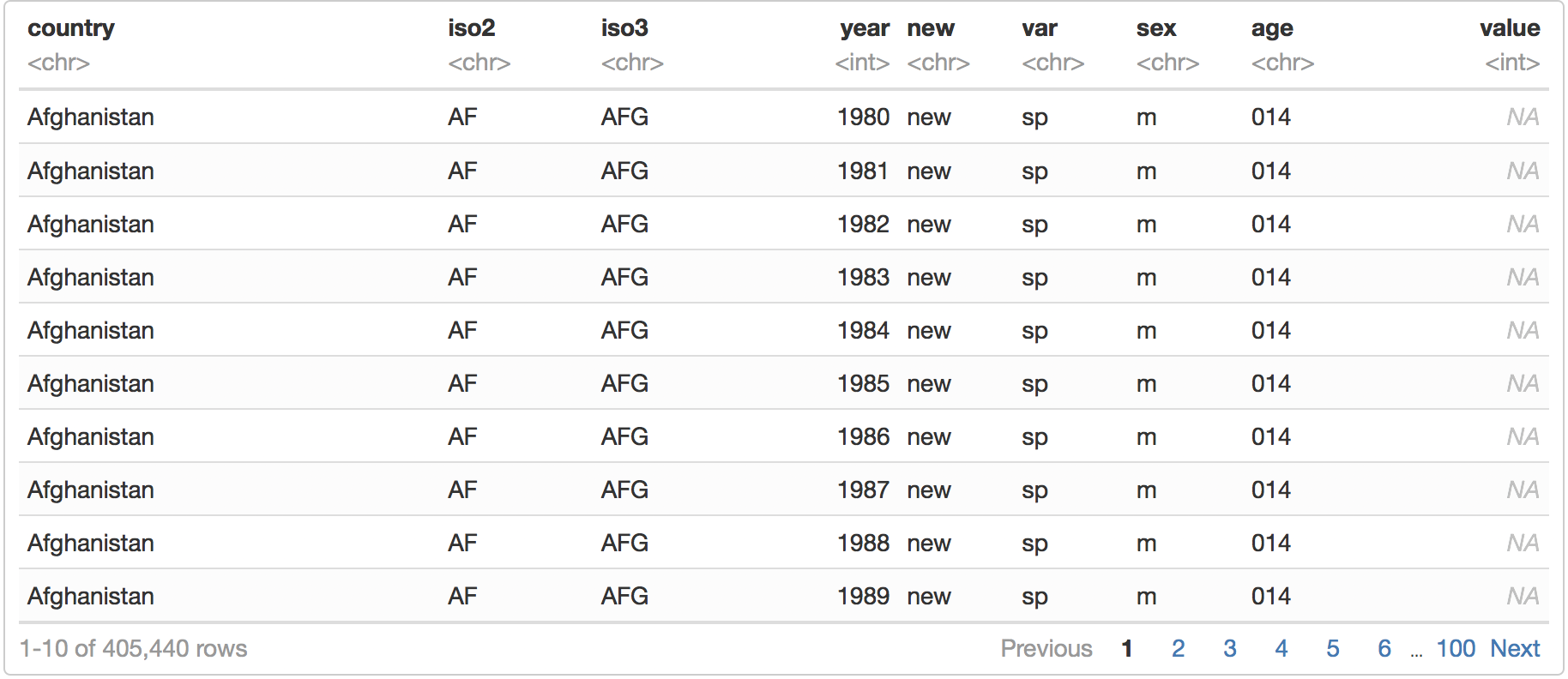
**1- Tidy Task 1:**

Use appropriate “tidyr” functions to reshape the WHO data set into the form given below:



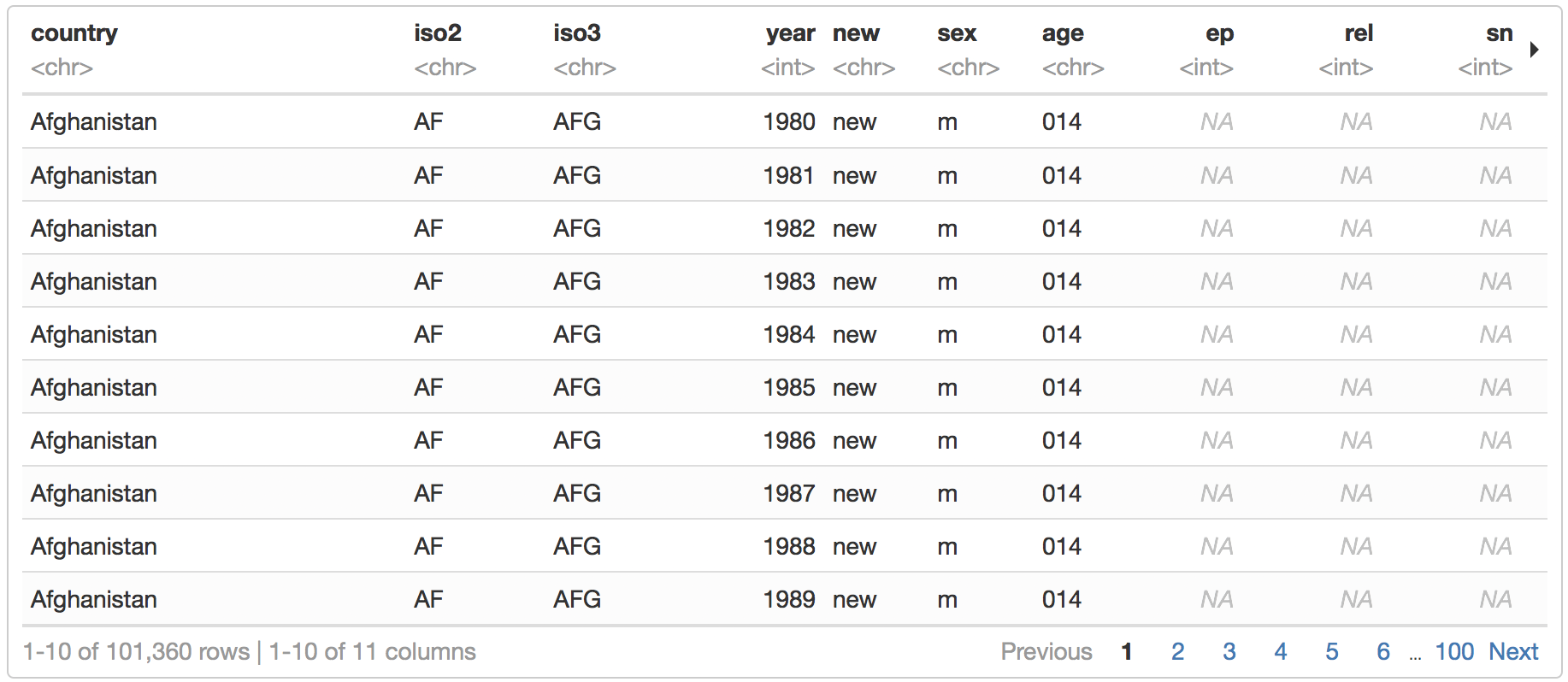
**2- Tidy Task 2:**

The WHO data set is not in a tidy format yet. The “code” column still contains four different variables' information (see variable description section for the details). Separate the “code” column and form four new variables using appropriate “tidyr” functions. The final format of the WHO data set for this task should be in the form given below:



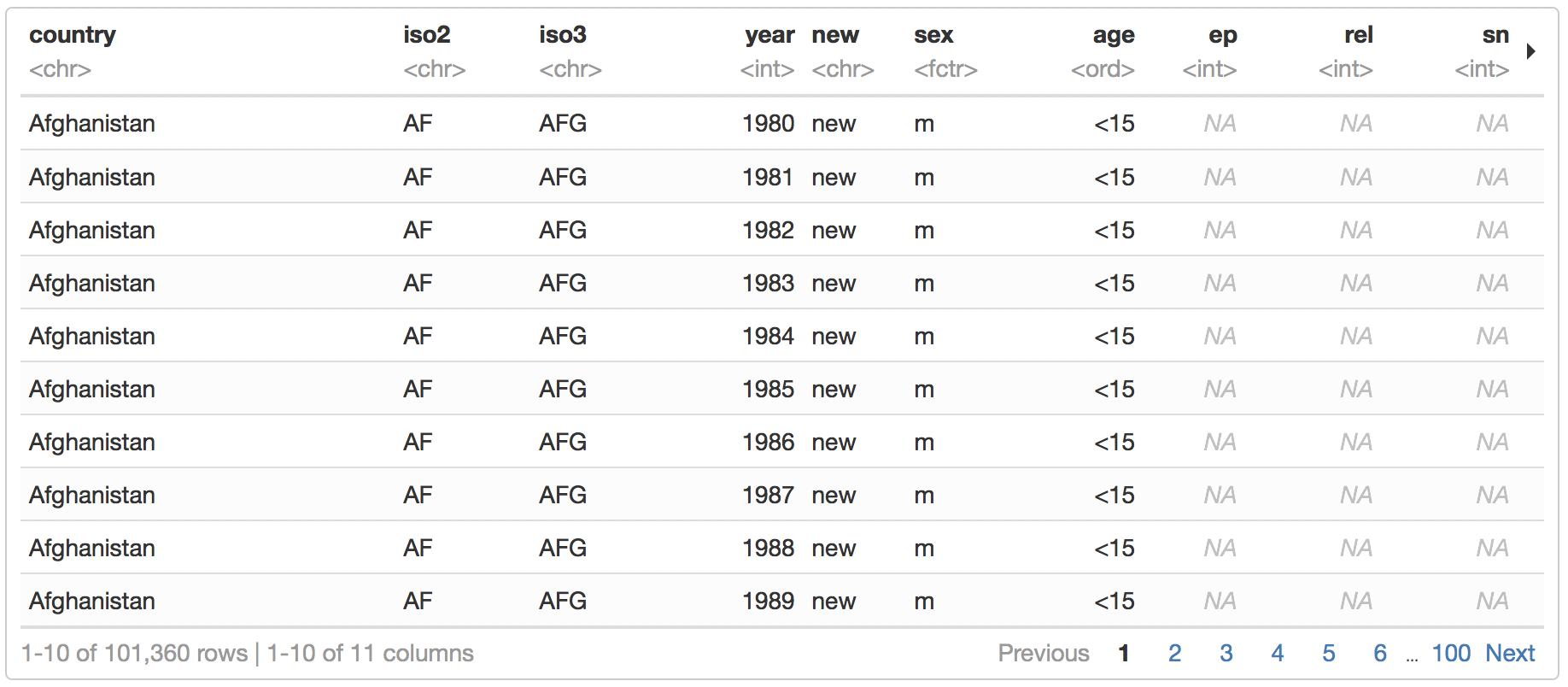
**3- Tidy Task 3:**

The WHO data set is not in a tidy format yet. The “rel”, “ep”, “sn”, and “sp” keys need to be in their own columns as we will treat each of these as a separate variable. In this step, move the “rel”, “ep”, “sn”, and “sp” keys into their own columns. The final format of the WHO data set for this task should be in the form given below:



**4- Tidy Task 4:**

There is one more step to tidy the WHO data set. We have two categorical variables “sex” and “age”. Use “mutate()” to factorise sex and age. For “age” variable, you need to create labels and also order the variable. Labels would be: <15, 15-24, 25-34, 35-44, 45-54, 55-64, 65>=. The final tidy version of the WHO data set would look like this:



**5- Task 5: Filter & Select**

Drop the redundant columns “iso2” and “new”, and filter any three countries from the tidy version of the WHO data set. Name this subset of the data frame as “WHO\_subset”.

You will use surveys and species data sets for Tasks 6 - 10. Read the species and surveys data sets using an appropriate function. Name these data frames as “species” and “surveys”, respectively.

**6- Task 6: Join**

Combine “surveys” and “species” data frames using the key variable “species\_id”. For this task, you need to add the species information (“genus”, “species”, “taxa”) to the “surveys” data. Rename the combined data frame as “surveys\_combined”.

**7- Task 7: Calculate**

Using the “surveys\_combined” data frame, calculate the average weight and hindfoot length of one of the species observed in each month (irrespective of the year). Make sure to exclude missing values while calculating the average.

**8- Task 8: Missing Values**

Select one of the years in the “surveys\_combined” dataframe, rename this data set as “surveys\_combined\_year”. Using “surveys\_combined\_year” dataframe, find the total missing values in “weight” column grouped by species. Replace the missing values in “weight” column with the mean values of each species. Save this imputed data as “surveys\_weight\_imputed”.

**9- Task 9: Inconsistencies or Special Values**

Inspect the “weight” column in “surveys\_weight\_imputed” dataframe for any further inconsistencies or special values (i.e., NaN, Inf, -Inf). Trace back and explain briefly why you got such a value.

**10- Task 10: Outliers**

Using the “surveys\_combined” data frame, inspect the variable hindfoot length for possible univariate outliers. If you detect any outliers use any of the methods outlined in the Module 6 notes to deal with them. Explain briefly the actions that you take to handle outliers.

Submission Instructions

The assignment 2 report must be completed using the R Markdown template provided here:

[R Markdown Template - Assignment 2](https://drive.google.com/open?id=1A21_MO0oemD0w3ZUdn-1jUkmjmMHSN4X)

You must use the headings and chunks provided in the template, you may add additional R chunks if you require. **In the report, all R chunks and outputs needs to be visible. Failure to do so will result in a loss of marks.**

The report must be uploaded to Turnitin as a **PDF** with your code chunks showing. The easiest way to achieve this is to **Preview** your notebook in HTML (by clicking Preview) → **Open in Browser** (Chrome) → Right click on the report in Chrome → Click **Print** and Select the **Destination** Option to **Save as PDF**.

**All group members must submit a copy of the report! Group members that are not registered and do not submit a report will not be acknowledged. One group member’s submission will be marked and given feedback. It will be the responsibility of the marked group member to share the group’s feedback with the other group members. The other group members will receive a mark only.**

Extensions will only be granted in accordance with the [RMIT University Extension and Special Consideration Policy](http://www1.rmit.edu.au/browse;ID=8r23xxpe4fne). No exceptions. Assignments submitted late will be penalised (see [Course Information](https://drive.google.com/open?id=177dgNO0PezpPfqEtXziUlkmVLxfq5dORDnxAb03YIRM) for further details).

Collaboration

You are permitted to discuss and collaborate on the assignment with your classmates. However, the write-up of the report must be an individual/group effort. Assignments will be submitted through Turnitin, so if you’ve copied from a fellow classmate/group, it will be detected. It is your responsibility to ensure you do not copy or do not allow another classmate/groups to copy your work. If plagiarism is detected, both the copier and the student/group copied from will be responsible. It is good practice to never share assignment files with other students/groups. You should ensure you understand your responsibilities by reading the RMIT University website on [academic integrity](http://www1.rmit.edu.au/browse;ID=kkc202lwe1yv). Ignorance is no excuse.

Learning Objectives Assessed

This assignment assesses the following Course Learning Objectives:

1. Critically reflect upon different data sources, types, formats and structures.

2. Apply data integration techniques to import and combine different sources of data.

3. Apply different data manipulation techniques to recode, filter, select, split, aggregate, and reshape the data into a format suitable for statistical analysis.

4. Justify data by detecting and handling missing values, outliers, inconsistencies and errors.

# Assignment 2 Marking Rubric

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Not acceptable**  **(0)** | **Needs Improvement**  **(1)** | **Excellent**  **(2)** |
| **Task 1**  **(10%)** | Unable to tidy the data set properly. | There was an attempt to tidy the data but it wasn’t in the required form | A complete set of tasks were provided to tidy the data set in the required form. |
| **Task 2 (10%)** | Unable to tidy the data set properly. | There was an attempt to tidy the data but it wasn’t in the required form | A complete set of tasks were provided to tidy the data set in the required form. |
| **Task 3 (10%)** | Unable to tidy the data set properly. | There was an attempt to tidy the data but it wasn’t in the required form | A complete set of tasks were provided to tidy the data set in the required form. |
| **Task 4 (10%)** | Unable to tidy the data set properly. | There was an attempt to tidy the data but it wasn’t in the required form | A complete set of tasks were provided to tidy the data set in the required form. |
| **Task 5 (10%)** | Unable to filter and select the required subset properly. | There was an attempt to filter the data set but unable to select the three countries. | A complete set of tasks were provided to filter and select the required subset of the data. |
| **Task 6 (10%)** | Unable to join two data sets. | Able to join the two data sets but there was a room for improvement (i.e. the output was missing) | A complete set of tasks were provided to join two data sets. |
| **Task 7 (10%)** | Unable to calculate the average weight and hindfoot length of one of the species within each month. | Able to calculate the average weight and hindfoot length of one of the species but there was a room for improvement (i.e. averages were not grouped by month, missing values weren’t excluded while calculating the average) | A complete set of tasks were provided to calculate the average weight and hindfoot length of one of the species within each month. |
| **Task 8 (10%)** | Unable to scan the weight variable for missing values and impute the missing values with the mean of each species | Able to scan the weight for missing values, but there was a room for improvement (i.e. total number of missing values were not reported grouped by species and/ or unable to impute the missing value(s) with the mean of each species. | A complete set of tasks were provided to scan and impute the missing values in the weight variable. |
| **Task 9 (10%)** | Unable to inspect weight variable for inconsistencies or special values. | Able to inspect weight variable for inconsistencies or special values, but there was a room for improvement (i.e. unable to provide an explanation for such inconsistencies/special values. | A complete set of tasks were provided to inspect the weight for inconsistencies or special values. |
| **Task 10 (10%)** | Unable to scan for and deal with outliers. | Able to scan for and deal with outliers, but there was a room for improvement (i.e. brief explanation of actions for handling outliers was missing) | A complete set of tasks were provided to scan for and deal with outliers. |