MATH2349 Data Preprocessing Assignment 2

(Last Updated 9.9.2019)

Weight: 10%
Due date: 22 September 2019, 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 10% and is due **22/09/2019**.

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 15/09/2019 to register their group details. After the deadline, group registrations won't be accepted.** Submit the details of your group here:

Group Registration Form

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). 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/) 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:

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new_sp_m014 <int>	new_sp_m1524 <int>	new_sp_m2534 <int>	new_sp_m3544 <int>	new_sp_m4554 <int>
Afghanistan	AF	AFG	1980	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1981	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1982	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1983	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1984	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1985	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1986	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1987	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1988	NA	NA	NA	NA	NA
Afghanistan	AF	AFG	1989	NA	NA	NA	NA	NA

1-10 of 7,240 rows | 1-9 of 60 columns Previous 1 2 3 4 5 6 … 100 Next

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) 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):

species_id <chr>	genus <chr>	species <chr>	taxa <chr>
AB	Amphispiza	bilineata	Bird
AH	Ammospermophilus	harrisi	Rodent
AS	Ammodramus	savannarum	Bird
BA	Baiomys	taylori	Rodent
CB	Campylorhynchus	brunneicapillus	Bird
CM	Calamospiza	melanocorys	Bird

Surveys data (first six observations):

record_id <int>	month <int>	day <int>	year <int>	species_id <chr>	sex <chr>	hindfoot_length <int>	weight <int>
1	7	16	1977	NL	M	32	*NA*
2	7	16	1977	NL	M	33	*NA*
3	7	16	1977	DM	F	37	*NA*
4	7	16	1977	DM	M	36	*NA*
5	7	16	1977	DM	M	35	*NA*
6	7	16	1977	PF	M	14	*NA*

6 rows

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:

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	code <chr>	value <int>
Afghanistan	AF	AFG	1980	new_sp_m014	NA
Afghanistan	AF	AFG	1981	new_sp_m014	NA
Afghanistan	AF	AFG	1982	new_sp_m014	NA
Afghanistan	AF	AFG	1983	new_sp_m014	NA
Afghanistan	AF	AFG	1984	new_sp_m014	NA
Afghanistan	AF	AFG	1985	new_sp_m014	NA
Afghanistan	AF	AFG	1986	new_sp_m014	NA
Afghanistan	AF	AFG	1987	new_sp_m014	NA
Afghanistan	AF	AFG	1988	new_sp_m014	NA
Afghanistan	AF	AFG	1989	new_sp_m014	NA

1-10 of 405,440 rows Previous 1 2 3 4 5 6 ... 100 Next

Show your R codes to reshape the data and provide the output of the final dataset along with the dimensions. Failure to do this would result in a reduction in the mark.

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 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:

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	var <chr>	sex <chr>	age <chr>	value <int>
Afghanistan	AF	AFG	1980	new	sp	m	014	NA
Afghanistan	AF	AFG	1981	new	sp	m	014	NA
Afghanistan	AF	AFG	1982	new	sp	m	014	NA
Afghanistan	AF	AFG	1983	new	sp	m	014	NA
Afghanistan	AF	AFG	1984	new	sp	m	014	NA
Afghanistan	AF	AFG	1985	new	sp	m	014	NA
Afghanistan	AF	AFG	1986	new	sp	m	014	NA
Afghanistan	AF	AFG	1987	new	sp	m	014	NA
Afghanistan	AF	AFG	1988	new	sp	m	014	NA
Afghanistan	AF	AFG	1989	new	sp	m	014	NA

1-10 of 405,440 rows Previous 1 2 3 4 5 6 ... 100 Next

Show your R codes to reshape the data and provide the output of the final dataset along with the dimensions. Failure to do this would result in a reduction in the mark.

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:

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	sex <chr>	age <chr>	ep <int>	rel <int>	sn <int>
Afghanistan	AF	AFG	1980	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1981	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1982	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1983	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1984	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1985	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1986	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1987	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1988	new	m	014	NA	NA	NA
Afghanistan	AF	AFG	1989	new	m	014	NA	NA	NA

1-10 of 101,360 rows | 1-10 of 11 columns Previous 1 2 3 4 5 6 ... 100 Next

Show your R codes to reshape the data and provide the output of the final dataset along with the dimensions. Failure to do this would result in a reduction in the mark.

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:

country <chr>	iso2 <chr>	iso3 <chr>	year <int>	new <chr>	sex <fctr>	age <ord>	ep <int>	rel <int>	sn <int>
Afghanistan	AF	AFG	1980	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1981	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1982	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1983	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1984	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1985	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1986	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1987	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1988	new	m	<15	NA	NA	NA
Afghanistan	AF	AFG	1989	new	m	<15	NA	NA	NA

1-10 of 101,360 rows | 1-10 of 11 columns Previous 1 2 3 4 5 6 ... 100 Next

Show your R codes to reshape the data and provide the output of the final dataset along with the dimensions. Failure to do this would result in a reduction in the mark.

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”. **Show**

your R codes, provide the output and the dimensions of the "WHO_subset". Failure to do this would result in a reduction in the mark.

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". **Show your R codes, provide the output and the dimensions of the "surveys_combined". Failure to do this would result in a reduction in the mark.**

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. **Show your R codes and provide the output for each step. Failure to do this would result in a reduction in the mark.**

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 the "weight" column grouped by species. Replace the missing values in the "weight" column with the mean values of each species. Save this imputed data as "surveys_weight_imputed". **Show your R codes, provide the outputs of each step and check whether the imputation was successful in the "surveys_weight_imputed" data. Failure to do this would result in a reduction in the mark.**

9- Task 9: Special Values

Inspect the "weight" column in "surveys_weight_imputed" dataframe for any special values (i.e., NaN, Inf, -Inf). If you detect any special values, trace back and explain briefly why you got such a value. **Show your R codes and provide the outputs for each step. Failure to do this would result in a reduction in the mark.**

10- Task 10: Outliers

Using the "surveys_combined" data frame, inspect the variable hindfoot length for possible univariate outliers. If you detect any univariate outliers, use a suitable method outlined in the

Module 6 notes to deal with them. Explain your methodology (i.e. explain why you have chosen that methodology and explain the actions that you have taken to handle outliers) and communicate your results clearly. **Show your R codes and provide the outputs for each step. Failure to do this would result in a reduction in the mark.**

Submission Instructions

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

R Markdown Template - Assignment 2

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. No exceptions. Assignments submitted late will be penalised (see Course Information 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. 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 (2)	Meets Expectations (3)	Excellent (4)
Task 1 (5%)	Unable to tidy the data set properly: Some R scripts were provided in an attempt to tidy the data set, however the scripts were not working properly (generating error message) and/or the final data wasn't in the required form (i.e. dimensions of the final data set didn't match).	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set matched, however the output of the final data was not given.	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set match and output of the final data was provided however, the task could be coded in a more efficient way.	A complete set of tasks were provided to tidy the data set in the required form: - Correct R scripts were provided - Dimension of the final data matched - The output of the final data was provided - The task was coded in an efficient way
Task 2 (5%)	Unable to tidy the data set properly: Some R scripts were provided in an attempt to tidy the data set, however the scripts were not working properly (generating error message) and/or the final data wasn't in the required form (i.e. dimensions of the final data set didn't match).	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set matched, however the output of the final data was not given.	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set match and output of the final data was provided however, the task could be coded in a more efficient way.	A complete set of tasks were provided to tidy the data set in the required form: - Correct R scripts were provided - Dimension of the final data matched - The output of the final data was provided - The task was coded in an efficient way
Task 3 (5%)	Unable to tidy the data set properly: Some R scripts were provided in an attempt to tidy the data set, however the scripts were not working properly (generating error message) and/or the final data wasn't in the required form (i.e. dimensions of the final data set didn't match).	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set matched, however the output of the final data was not given.	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set match and output of the final data was provided however, the task could be coded in a more efficient way.	A complete set of tasks were provided to tidy the data set in the required form: - Correct R scripts were provided - Dimension of the final data matched - The output of the final data was provided - The task was coded in an efficient way
Task 4 (5%)	Unable to tidy the data set properly: Some R scripts were provided in an attempt to tidy the data set, however the scripts were not working properly (generating error message) and/or the final data wasn't in the required form (i.e. dimensions of the final data set didn't match).	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set matched, however factorisation of the age and/or sex variables needed improvement or the output of the final data was not given.	Correct R scripts were provided to tidy the data set in the required form, the dimensions of the final data set matched, age and sex variables were factorised properly and the output was provided, however the task could be coded in a more efficient way.	A complete set of tasks were provided to tidy the data set in the required form: - Correct R scripts were provided - Dimension of the final data matched - Age and sex variables were factorised properly - The output of the final data was provided - The task was coded in an efficient way
Task 5 (5%)	Unable to retrieve the required subset properly: Some R scripts were provided in an attempt to get the required subset, however the scripts were not working properly (generating error message) and/or didn't retrieve the correct subset.	There was an attempt to get the required subset of the data but one of the tasks needed improvement. For example; - Unable to select the three countries, OR - Unable to drop redundant columns, OR	Correct R scripts were provided to retrieve the required subset and the output of the final subset was provided however, the task could be coded in a more efficient way.	A complete set of tasks were provided to retrieve the required subset of the data: - Correct R scripts were provided to select three countries and drop redundant columns - The output of the final subset was provided - The task was coded in an efficient way

		- The output of the final subset wasn't provided		
Task 6 (10%)	Unable to join surveys and species data frames by adding the species information (“genus”, “species”, “taxa”) to the surveys data: Some R scripts were provided in attempt to join data frames however scripts were not working properly (generating error message).	There was an attempt to join the data frames but one of the tasks needed improvement. For example; - An incorrect join function was used, OR - The output of the merged data was missing	Able to join the two data frames by adding the species information (“genus”, “species”, “taxa”) to the “surveys” data. Correct R scripts were provided and the output of the merged data was given, however, the task could be coded in a more efficient way.	A complete set of tasks were provided to join data frames: - Correct R scripts were provided to join the two data frames by adding the species information (“genus”, “species”, “taxa”) to the surveys data. - The output of the merged data was provided - The task was coded in an efficient way
Task 7 (15%)	Unable to calculate the average weight and hindfoot length of one of the species within each month. Some scripts were provided in attempt to do this however the scripts were not working properly (generating error message) or not producing the correct averages due to one of the following: - One of the species was not selected OR - Averages were not grouped by month	There was an attempt to calculate the average weight and hindfoot length of one of the species but the task heavily needed improvements in at least one of the following: - Missing values weren’t excluded while calculating the averages - One of the average calculations was missing - Output of the averages per month was missing	A complete set of tasks were provided to calculate the average weight and hindfoot length of one of the species within each month. Therefore the student was able to select one of the species, calculate the averages grouped by month, calculate the averages by excluding missing values, calculate all of the averages, provide the output for the averages per month. However, the task could be coded in a more efficient way.	A complete set of tasks were provided to calculate the average weight and hindfoot length of one of the species within each month: Student was able to: - Select one of the species - Calculate averages grouped by month - Exclude missing values while calculating the averages - Calculate averages for both weight and hindfoot length - Provide the output of the averages per month - Provide an efficient and clever way of coding.
Task 8 (20%)	Unable to scan the weight variable for missing values and impute the missing values with the mean of each species. Some scripts were provided in attempt to do this however the scripts were not working properly (generating error message) or unable to impute missing values with the correct mean values due to one of the following: - Unable to select one of the years, OR - The total number of missing values grouped by species were not reported	There was an attempt to scan the weight for missing values, but imputing part heavily needed improvements because the student was: - Unable to impute the missing value(s) with the mean of each species.	Able to scan the weight variable for missing values and impute the missing values with the mean of each species. However the result of the imputation was not checked whether it is successful or not.	A complete set of tasks were provided to scan and impute the missing values in the weight variable. Student was able to: - Select one of the years - Report the total number of missing values grouped by species - Impute the missing value(s) with the mean of each species. - Check whether the imputation was successful
Task 9 (10%)	There was no attempt to inspect weight variable for special values (i.e., NaN, Inf, -Inf).	There was an attempt to inspect weight variable for special values, but the task heavily needed improvements in one of the following: - Unable to check NaN values, OR - Unable to check Inf, -Inf values, OR - Unable to trace back and provide an explanation for such special values.	Able to inspect the weight for special values (i.e., NaN, Inf, -Inf), trace back and provide an explanation for such special values, however the task could be coded or the outputs could be presented in a more efficient way.	A complete set of tasks were provided: - Weight variable was checked for special values. - Trace back and provide an explanation for such special values - Provide an efficient and clever way of coding and presenting outputs

Task 10 (20%)	Unable to scan for and deal with outliers. Some scripts were provided in attempt to scan for outliers in the data however no methodology/actions were taken to handle those outliers.	Able to scan for and deal with outliers, but the task heavily needed improvements in the methodology. For example: - A methodology was applied to scan and deal with outliers however there was no attempt to check whether this approach can be safely applied, OR - A methodology was applied to scan and deal with outliers however the approach taken was not suitable/safe to apply.	Able to follow a safe and suitable methodology to scan and deal with outliers, however the methodology was not explained enough or the results and outputs weren't presented in a clearer way.	A complete set of tasks were provided to scan for and deal with outliers. - A safe and suitable methodology was followed to scan and deal with outliers - The methodology taken was explained thoroughly. - The results and outputs were presented clearly.