Sample questions

DSA2101 AY22/23 Semester 2

Here are some sample questions to help guide your preparation for the midterm test.

Solve them by yourself first. We will go over the solutions on Friday of Week 7.

Instructions

1. Please answer all questions in a single `R` Markdown file.
2. The exam data files will be available on Canvas *five minutes* before the exam. Please arrive early on the exam day for necessary setups.
3. The packages you need are `readxl`, `stringr`, `lubridate`, `tidyverse`. Ensure you have installed them *prior to* the exam. You won't have internet access during the exam.
4. Interpret each question as best you can, on your own, to your best judgement.
5. Before you submit your `R` Markdown file, ensure that your `Rmd` can knit to HTML (i.e., no syntax errors or crashes).
6. Submit your code to **both** Examplify and Canvas

   1) Copy and paste your entire `Rmd` code to the Examplify text box.
   2) Save an **exact copy** of the `Rmd` file on your laptop for submitting on Canvas **immediately after** the exam. Do not modify your code in the submission file. Any difference between Examplify and Canvas submissions will be penalized.

Please answer all questions in a single `R` Markdown file, and upload the file to both Examplify and Canvas at the end of the exam.

Question 1. UNICEF HIV data

In this section, we shall work with the data file `UNICEF_hiv_data.xlsx`. The data are obtained from UNICEF's Data Warehouse. The data are formatted as follows:

- The first column is the country name.
- The next two columns contain information on the indicators.
- The rest contain statistics on child HIV infections and deaths from 2010 to 2021.
- A dash `-` indicates that no data is available.
- An analyst tells you that `<0.01` means that the value is negligible or unreliable.

Your task is to turn this into a tidy data table for further analysis:

1. Describe how would a tidy data table look like. What columns a tidy layout of this data set would have and why. Write a brief summary (in 50 – 100 words) in an `Rmd` text section entitled "Question 1.1".

2. Convert the data into a tidy data layout. Replace all `-` and `<0.01` with `NA`s and convert all estimates into numeric. Save the resulting data frame as `unicef_hiv_tidy`.

3. Create a graph to display the global evolution of HIV incidence rate for children aged 0-14 years from 2010 to 2021.

4. Suppose we want to further analyze the geographical disparity in HIV incidence rates. To do so, we need additional data on the geographical location of these countries.

   The file `iso_unicef_region.csv` contain information on country name, country code, and the UNICEF region name. Use the file and create a data visualization that compares the HIV incidence rates in Eastern and Southern Africa to the global average across years.

   For this question, you may ignore the countries/regions that do not link to any UNICEF region.

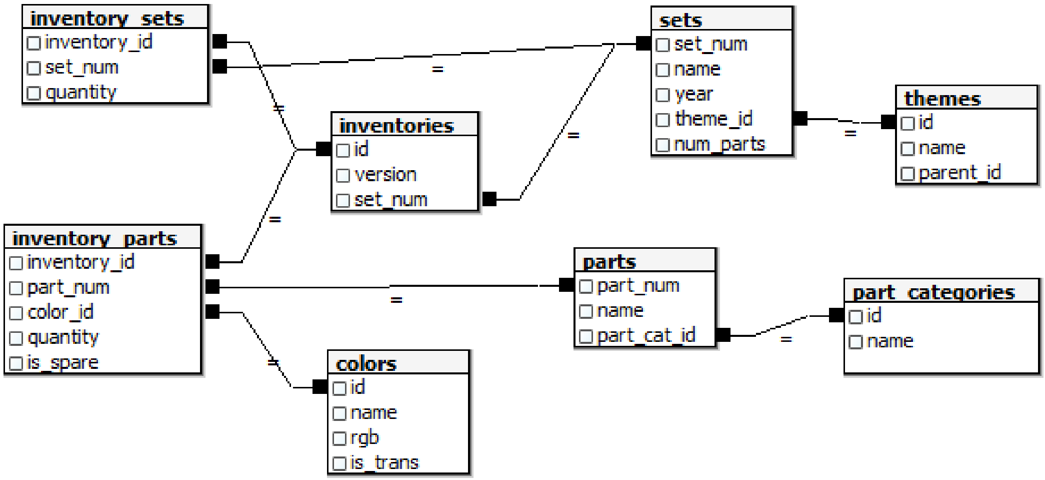
Question 2. LEGO data

The data set we shall work on for this question is on LEGO bricks. LEGO is a brand of toy building bricks. Bricks are often sold in sets, which allow the owner to build specific objects. Each set can be classified under a theme. Each set contains parts, each of which can be identified by a part number and a part category. Parts also differ in terms of their color.

Download the data sets from Kaggle.

https://www.kaggle.com/datasets/rtatman/lego-database

Unzip the downloaded file and save the data sets under your `data` folder. There are eight CSV files in total. The data contain the LEGO parts/sets/colors and inventories of every official LEGO set in the Rebrickable database, as current as 2017. Here is a diagram that identifies the keys that link the files together. In database jargon, such diagram is called a schema.



Let us first read in all 8 data files and load the `tidyverse` package.

1. Consider the tables in `themes.csv` and `sets.csv`. How many **themes** do not have any associated **sets**?

2. Create a table that contains the **color** for each **part number** associated with a **part category** that contains the word `Bricks`. Store the table as `brick_parts`. The table should contain at least the following information.

   - part number
   - part name
   - part category name
   - color name
   - color rgb code

The first three rows of the table should look like:

```
##   part_num                                                         part_name
## 1    10314 Brick 1 x 4 x 1 1/3 No Studs, Curved Top, with Raised Inside Support
## 2    10314 Brick 1 x 4 x 1 1/3 No Studs, Curved Top, with Raised Inside Support
## 3    10314 Brick 1 x 4 x 1 1/3 No Studs, Curved Top, with Raised Inside Support
##   part_category_name color_name    rgb
## 1      Bricks Curved      Black 05131D
## 2      Bricks Curved      Black 05131D
## 3      Bricks Curved     Orange FE8A18
```

3. Use the **themes** table to identify all themes associated with the Star Wars franchise. Create a table named `pop_themes` that contains the information below.

   - the number of sets released each year
   - the average number of parts per set each year

   The first few rows of the table should look like:

```
## # A tibble: 5 x 3
##    year num_sets avg_num_parts
##   <int>    <int>         <dbl>
## 1  1999       14          268.
## 2  2000       27          234.
## 3  2001       14          390.
## 4  2002       30          285.
## 5  2003       20          417.
```

4. When we perform a left join between the `sets` and `inventories` tables, the number of rows changes. Why? Identify the cause for this. Discuss (in 50 words) what you found in an `Rmd` text section entitled "Question 2.4".

5. Explore the data set and come up with one question you find interesting about the data. Create a graph that answers the question. Include the code you use. Summarize (in 50 – 100 words) your findings in an `Rmd` text section entitled "Question 2.5".

   Here are some prompts to help frame your discussion:

   - How would you answer the question with this data? What summary statistics and visualization helped you answer this question?
   - Create a visualization from this data which confirms something you thought would be true. Why did you expect to see this?
   - Create a visualization from this data that you did not expect to see, or were surprised to see. Why was this surprising?

Requirements

- Your code must generate the objects required by each question.
  - Data frames (tibbles): `unicef_hiv_tidy`, `brick_parts`, `pop_themes`
- Three required plots. You only need to use base `R` plotting. But if you are already comfortable with `ggplot`, you may also go ahead and use it. Remember to include axis labels and necessary legends.
- There must also be two short discussion sections, under `Rmd` text sections entitled "Question 1.1", "Question 2.4", and "Question"2.5", respectively.