Programming Assignment: Time Usage

You have not submitted. You must earn 8/10 points to pass.

Deadline Pass this assignment by July 2, 11:59 PM PDT

Instructions My submission Discussions

Note: If you have paid for the Certificate, please make sure you are submitting to the required assessment and not the optional assessment. If you mistakenly use the token from the wrong assignment, your grades will not appear

To start, first download the assignment: timeusage.zip. For this assignment, you also need to download the data (164 MB):

http://alaska.epfl.ch/~dockermoocs/bigdata/atussum.csv

and place it in the folder src/main/resources/timeusage/ in your project directory.

The problem

The dataset is provided by Kaggle and is documented here:

https://www.kaggle.com/bls/american-time-use-survey

It contains information about how do people spend their time (e.g., sleeping, eating, working, etc.).

Here are the first lines of the dataset:

Our goal is to identify three groups of activities:

- primary needs (sleeping and eating),
- work,
- other (leisure).

And then to observe how do people allocate their time between these three kinds of activities, and if we can see differences between men and women, employed and unemployed people, and young (less than 22 years old), active (between 22 and 55 years old) and elder people.

At the end of the assignment we will be able to answer the following questions based on the dataset:

- how much time do we spend on primary needs compared to other activities?
- do women and men spend the same amount of time in working?
- does the time spent on primary needs change when people get older?
- how much time do employed people spend on leisure compared to unemployed people?

To achieve this, we will first read the dataset with Spark, transform it into an intermediate dataset which will be easier to work with for our use case, and finally compute information that will answer the above questions.

Read-in Data

The simplest way to create a DataFrame consists in reading a file and letting Spark-sql infer the underlying schema. However this approach does not work well with CSV files, because the inferred column types are always String.

In our case, the first column contains a String value identifying the respondent but all the other columns contain numeric values. Since this schema will not be correctly inferred by Spark-sql, we will define it programmatically. However, the number of columns is huge. So, instead of manually enumerating all the columns we can rely on the fact that, in the CSV file, the first line contains the name of all the columns of the dataset.

Our first task consists in turning this first line into a Spark-sql StructType. This is the purpose of the dfSchema method. This method returns a StructType describing the schema of the CSV file, where the first column has type StringType and all the others have type DoubleType. None of these columns are nullable.

The second step is to be able to effectively read the CSV file is to turn each line into a Spark-sql Row containing columns that matches the schema returned by dfSchema. That's the job of the row method.

Project

As you probably noticed, the initial dataset contains lots of information that we don't need to answer our questions, and even the columns that contain useful information are too much detailed. For instance, we are not interested in the exact age of each respondent, but just whether she was "young", "active" or "elder".

Also, the time spent on each activity is very detailed (there are more than 50 reported activities). Again, we don't need this level of detail, we are only interested in three activities: primary needs, work and other.

So, with this initial dataset it would a bit hard to express the queries that would give us the answers we are looking for.

The second part of this assignment consists in transforming the initial dataset into a format that will be easier to work with.

A first step in this direction is to identify which columns are related to the same activity. Based on the description of the activity corresponding to each column (given in this document), we deduce the following rules:

- "primary needs" activities (sleeping, eating, etc.) are reported in columns starting with "t01", "t03", "t11", "t1801" and "t1803" ;
- working activities are reported in columns starting with "t05" and "t1805" ;

- other activities (leisure) are reported in columns starting with "t02", "t04", "t06", "t07", "t08", "t09", "t10", "t12", "t13", "t14", "t15", "t16" and "t18" (only those which are not part of the previous groups).

Then our work consists in implementing the classifiedColumns method, which classifies the given list of column names into three groups (primary needs, work or other). This method should return a triplet containing the "primary needs" columns list, the "work" columns list and the "other" columns list.

The second step is to implement the timeUsageSummary method, which projects the detailed dataset into a summarized dataset. This summary will contain only 6 columns: the working status of the respondent, his sex, his age, the amount of daily hours spent on primary needs activities, the amount of daily hours spent on working and the amount of daily hours spent on other activities.

Each "activity column" will contain the sum of the columns related to the same activity of the initial dataset. Note that time amounts are given in minutes in the initial dataset, whereas in our resulting dataset we want them to be in hours.

The columns describing the work status, the sex and the age, will contain simplified information compared to the initial dataset.

Last, people that are not employable will be filtered out of the resulting dataset.

The comment on top of the timeUsageSummary method will give you more specific information about what is expected in each column.

Aggregate

Finally, we want to compare the *average time* spent on each activity, for all the combinations of working status, sex and age.

We will implement the timeUsageGrouped method which computes the average number of hours spent on each activity, grouped by working status (employed or unemployed), sex and age (young, active or elder), and also ordered by working status, sex and age. The values will be rounded with a scale of 1.

Now you can run the project and see what the final DataFrame contains. What do you see when you compare elderly men versus elderly women's time usage? How much time elder people allocate to leisure compared to active people? How much time do active employed people spend to work?

Alternative ways to manipulate data

We can also implement the timeUsageGrouped method by using a plain SQL query instead of the DataFrame API. Note that sometimes using the programmatic API to build queries is a lot easier than writing a plain SQL query. Can you think of a previous query that would have been a nightmare to write in plain SQL?

4

Finally, in the last part of this assignment we will explore yet another alternative way to express queries: using typed Datasets instead of untyped DataFrames.

5

Implement the timeUsageSummaryTyped method to convert a DataFrame returned by timeUsageSummary into a Dataset[TimeUsageRow]. The TimeUsageRow is a data type that models the content of a row of a summarized dataset. To achieve the conversion you might want to use the getAs method of Row. This method retrieves a named column of the row and attempts to cast its value to a given type.

6

Then, implement the timeUsageGroupedTyped method that performs the same query as timeUsageGrouped but uses typed APIs as much as possible. Note that not all the operations have a typed equivalent. round is an example of operation that has no typed equivalent: it will return a Column that you will have to turn into a TypedColumn by calling .as[Double]. Another example is orderBy, which has no typed equivalent. So for this operation to work be sure that your Dataset has a schema (because it relies on named columns, and column names are generally lost when using typed transformations).

How to submit

Copy the token below and run the submission script included in the assignment download. When prompted, use your email address **jeff_brown2@msn.com**.

Generate new token

Your submission token is unique to you and should not be shared with anyone.

You may submit as many times as you like.

Remaining pages removed

Alternative ways to manipulate data

We can also implement the timeUsageGrouped method by using a plain SQL query instead of the DataFrame API. Note that sometimes using the programmatic API to build queries is a lot easier than writing a plain SQL query. Can you think of a previous query that would have been a nightmare to write in plain SQL?

Finally, in the last part of this assignment we will explore yet another alternative way to express queries: using typed Datasets instead of untyped DataFrames.

Implement the timeUsageSummaryTyped method to convert a DataFrame returned by timeUsageSummary into a DataSet[TimeUsageRow]. The TimeUsageRow is a data type that models the content of a row of a summarized dataset. To achieve the conversion you might want to use the getAs method of Row. This method retrieves a named column of the row and attempts to cast its value to a given type.

Then, implement the timeUsageGroupedTyped method that performs the same query as timeUsageGrouped but uses typed APIs as much as possible. Note that not all the operations have a typed equivalent. round is an example of operation that has no typed equivalent: it will return a Column that you will have to turn into a TypedColumn by calling .as[Double]. Another example is orderBy, which has no typed equivalent. So, for this operation to work be sure that your Dataset has a schema (because it relies on named columns, and column names are generally lost when using typed transformations).