DSA2101 mid-term AY20/21 Sem 1

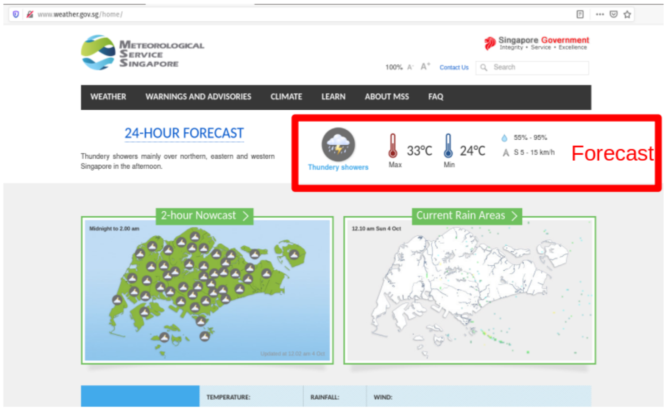
Contents

Instructions

1. Please answer all questions in a single R markdown file.
2. This exam is to be completed **individually**. Anyone found to have colluded will be given a 0, and will be reported.
3. The total marks allocation is 25 marks.
4. Do not post messages about the exam on the MSteams chat group.
5. Interpret each question as best you can, on your own, to your best judgement. Be sure to check your dataset for "dirty" data (not untidy data) at each step. It is ok to have some manual steps in your cleaning, but try to be as efficient as possible.
6. If a particular problem is difficult, don't give up completely - try to get as far as you can. Manage your time wisely.
7. Ensure that your Rmd file can knit to html before you submit. The solution checker is not available during the exam. Remember to use the usual relative path settings!
8. Submit your R markdown file to LumiNUS before 9pm Monday 12th October 2020.
9. Good luck!

Question 1 (5 marks)

The website http://www.weather.gov.sg/home contains a forecast of weather conditions in Singapore over the next 24 hours. This information can be found in the top right-hand corner of that website (see the red rectangle labelled "Forecast" below):



Thundery showers mainly over northern, eastern and western Singapore in the afternoon.

Write a function called `get_forecast` with *no arguments*, that will extract this forecast from that website. It would return a dataframe with 9 columns and 1 row:

```
       fcast_datetime        description min_temp max_temp min_humid max_humid
1 2020-10-10 23:52:50 Heavy thundery ...       23       32        65        95

wind_dir min_wind max_wind
      SS       15       25
```

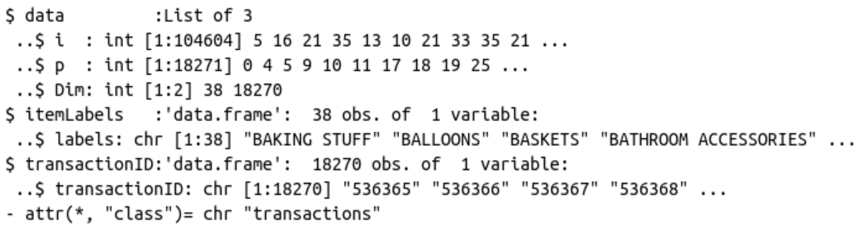
Question 2 (12 marks)

The dataset `wine_reviews_midterm_2021.xlsx` contains wine reviews from a magazine. This is a **modified version** of the dataset from https://www.kaggle.com/zynicide/wine-reviews. If you need more information, you may refer to that website.

1. Read the data into R. Did you observe any warnings? Investigate why they are there and fix the issues **within R**. Save your cleaned dataset as `wine_mag_clean`.
2. Most entries in the `title` column contain the vintage (year of the wine), year of the winery, and the region from which the wine came (in parentheses). Extract the region and vintage information and store them in two new columns, called `region` and `vintage`. You may ignore the currently existing `region_1` and `region_2` columns. *Hint: Vintage should not be earlier than the 20th century.*
3. Convert the following columns to lower case:
    1. `title`
    2. `variety`
    3. `winery`
    4. `description`
4. Convert the `price` and `points` columns to numeric.
5. Explore the dataset, and answer one question you find interesting about this data. Include the code you used, and summarise (in words) what you tried.

Question 3 (8 marks)

The file `transactions.rds` contains transactions from sales through an online gift shop. The sales are stored in an S3 object of class `transactions`:



```
$ data        :List of 3
 ..$ i  : int [1:104604] 5 16 21 35 13 10 21 33 35 21 ...
 ..$ p  : int [1:18271] 0 4 5 9 10 11 17 18 19 25 ...
 ..$ Dim: int [1:2] 38 18270
$ itemLabels   :'data.frame':  38 obs. of  1 variable:
 ..$ labels: chr [1:38] "BAKING STUFF" "BALLOONS" "BASKETS" "BATHROOM ACCESSORIES" ...
$ transactionID:'data.frame':  18270 obs. of  1 variable:
 ..$ transactionID: chr [1:18270] "536365" "536366" "536367" "536368" ...
- attr(*, "class")= chr "transactions"
```

There were 38 categories of items sold at the store. Their names can be found in the `itemLabels` component.

The object contains information on 18270 invoices. The invoice numbers are stored in the `transactionID` component of the object.

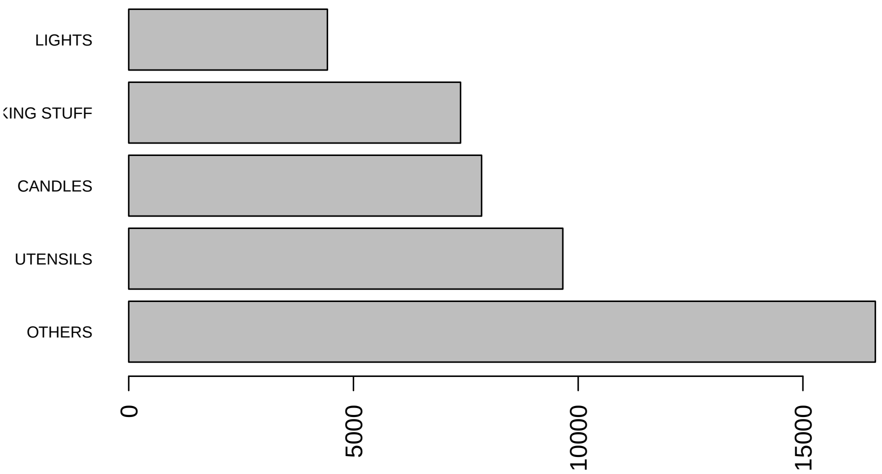
The items corresponding to each invoice are stored in a 38 by 18270 matrix of 1's and 0's. A one in row (i, j) of this matrix would indicate category i item was present in invoice j, where $1 \leq i \leq 38$ and $1 \leq j \leq 18270$. However, to save space, a sparse representation of this matrix is used - only the non-zero elements are stored. Component `p` is used to indicate which range of elements in `i` correspond to each of the 18270 invoices. For instance, it shows that 4 - 0 = 4 categories were present in the first invoice: categories 5, 16, 21 and 35. Similarly, 5 - 4 = 1 category was present in the second invoice: category 13.

1. Write a `plot` method for this class of objects, that displays the count of the `topN` item categories. Your method should call the `barplot` function from base R; it should not use ggplot. The signature of the argument should be:

```
function (x, topN = 10, ...)
```

(*Read up on the* `...` *argument on your own.*) Here is an example of the call and output. `trans_data` is an object of class `transactions`.

```r
plot(trans_data, topN=5, horiz=TRUE, las=2, cex.names=0.7)
```



2. The indexing operator that we use in R is just another function. Take a look:

```r
X <- 6:10
X[c(1,4,2)]
```

```
[1] 6 9 7
```

```
`[`(X, c(1,4,2))
```

```
[1] 6 9 7
```

Write an indexing method for objects of class `transactions`, that accesses invoice numbers. It would work this way:

```
trans_data[1:2]
```

```
  invoice        item
1  536365     CANDLES
2  536365      LIGHTS
3  536365      OTHERS
4  536365    UTENSILS
5  536366 HAND WARMER
```

```
trans_data[c(2, 4)]
```

```
  invoice        item
1  536366 HAND WARMER
2  536368      OTHERS
```

Requirements

1. A function named `get_forecast` for question 1.
2. A tibble/dataframe named `wine_mag_clean` for question 2.
3. Two S3 methods for objects of class `transactions` for question 3.