Agron5050 - Assignment 1

Version: 2 (2021 Oct 18)
Due: Week 7 2021/11/05 11:59pm

Download `Assignment1_data.zip` and use these files to answer the following questions.

- Please compiled all answers into a PDF report.
- Please show **ALL** your work, including the R codes, and regulation expressions you used. If you only present the final answer without any intermediate steps, you will receive **ZERO (0)** mark for that question.
- If you encounter issues while importing files into R, please have a look at GUIDE_Change_directory_in_RStudio.pdf

Question 1: Credit card number
You are developing an online shopping platform, and try to implement a simple credit card validation system. Here are the rules for three major credit card companies.

- VISA Card: Visa card numbers start with a 4. Total of 16 digits.
- MasterCard: MasterCard numbers start with the numbers 51 through 55. Total of 16 digits.
- Discover: Discover card numbers start with numbers 6011 or 6522, and end with numbers 2,4,6,8. Total of 16 digits.

Use the dataset `data_credit_card_number.txt` to determine how many valid credit card numbers are from each of the credit card companies. You can use a text editor or R to solve this. (3 marks)

Note: These numbers are NOT real credit card numbers. Please do NOT use them.

Question 2: Phone number
To expand your online shopping platform, therefore you are running lotteries for customers using their phone numbers. You notice that customers record their phone numbers in three different formats.

XXXXXXXXXX
XXXX-XXXXXX
XXXX-XXX-XXX

Use R and regular expressions to solve the following questions. You can import the data into R by using the following command.

```
> data <- readLines("data_phone_number.txt")
```

1. The telecommunication company Telecom have their mobile numbers begin with 0965, 0966, or 0967. How many customers are using mobile numbers from Telecom? (1 mark)
2. For round 1 of the lottery, you want to select customers with the 2nd to the last digits equals to 7 or 9. (i.e. XXXX-XXX-X9X and XXXX-XXX-X7X). How many customers win the round 1 lottery. (2 marks)
3. For round 2 of the lottery, you want to select customers with the 7th and 8th digits equal to 5. (i.e. XXXX-XX5-5XX). How many customers win the round 2 lottery. (2 marks)
4. Convert all phone numbers into this pattern XXXX-XXX-XXX. (Hints: use `gsub` in R) (3 marks)

Note: These numbers are NOT real phone numbers. Please do NOT call them.

Question 3: Shopping list

You receive a catalogue for some basic lab equipment from a biotech company. The catalogue file has two columns, catalogue_number and price. With a limited budget and you try to spend every single dollar, therefore, you have some constraints when you try to place an order.

Use R and regular expressions to find the catalogue number that satisfies the following scenarios. You can import the data into R by using the following command.

```
> data <- read.table("data_biotech_catalogue.tsv", header=T)
```

Your dataset should looks like this

```
> str(data)
'data.frame':   113 obs. of  2 variables:
 $ Catalogue_number: chr .....
 $ Price           : chr .....
```

Find all the catalogue_number that meet the criteria in each of the following scenarios.

- Scenario 1: Items that are less than (not include) $10. (1 mark)
- Scenario 2: Items with a price ending in the 22, 33, 55, 77 cents. (2 mark)
- Scenario 3: Items between $50 - $70, and the price ends with an even number (i.e. 55.32, 66.64, 51.98...etc) (2 marks)
- Scenario 4: Items between $50-$70, and catalogue_number begin with R (reagent). (3 marks)

Question 4: Sequencing primers

You try to design a primer to target the 16S rRNA for some specific bacteria. Use R and regular expressions to solve the following questions. Please use Table 1: IUPAC ambiguous codes for codes like **Y** and **M**. You can import the data into R by using the following command.

```
> data <- readLines("data_16S_rRNA.fasta")
```

The structure of the file: Each line starts with “>” Symbol, follows by sequence ID (i.e. SEQ_529), follows by a tab character, finally the rRNA sequence. Here is the snapshot of the first sequence.

```
> substring(data[1], 0, 30)
[1] ">SEQ_529\tAGAGUUUGAUCCUGGCUCAGG"
```

1. You start with the common primer 515F, which targets the following region. How many bacteria can be targeted by this primer? (2 marks)

   GUGYCAGCMGCCGCGGUAA

2. After you examine the results, you conclude that this primer doesn’t target the group of bacteria you want. Therefore, your labmate who work on the same group of bacteria give you a custom made primer set, which matches all the following

   AUAKWNNNWUGGG

   AUAKWNNNNNWUGGG

   How many bacteria can be targeted by this set of primers? (2 marks)

3. After examining the results, this set of primers is very close to what you are looking for. You further refine this primer set so it only targets the following 3 combinations.

   AUAGAUGAUUGGG

   AUAUAAGACCUUGGG

   AUAUUGGACAAUGGG

   Use one regular expression, and without using . (dot matches everything), and find the sequence ID (beginning of each line) associated with these three primers only. (4 marks)

   Version 2 amendment: Typo in these primers, highlight in orange.

Bonus Question (0 mark):

This is a challenge question worth zero (0) mark. You do not need to do it, nor to hand it in. Download “`assignment1_bonus_data.zip`”. This is an annotated E. Coli genome download from NCBI. In this file, you will find some genes like the following.

```
     CDS             1052..2152
                     /gene="dnaN"
                     /locus_tag="C7A06_RS00010"
                     /EC_number="2.7.7.7"
```

Try to find all genes with their associated EC_number (Enzyme Commission number), and reformat them as “`Gene:dnaN  EC_number:2.7.7.7`”

Not all genes are associated with EC_number, and not all EC_numbers are associated with a gene. Try to identify these as well.

Table 1: IUPAC Ambiguity Codes

Note: We are working with rRNA, therefore we will observe U (Uracil) instead of T (thymine).

Symbol	Description	Bases represented
A	Adenine	A – – –
C	Cytosine	– C – –
G	Guanine	– – G –
T or U	Thymine or Uracil	– – – T or – – – U
W	Weak	A – – T
S	Strong	– C G –
M	aMino	A C – –
K	Keto	– – G T
R	puRine	A – G –
Y	pYrimidine	– C – T
B	not A	– C G T
D	not C	A – G T
H	not G	A C – T
V	not T	A C G –
N	any Nucleotide	A C G T