RIntro Exam – 1st term 2023/24 – group 6pm

by Maria Kubara

Organisational information: The main purpose of this exam is to show how much you have learned from the RIntro course - or how to code in R. It doesn't really matter whether you use base functions or tidyverse solutions to solve these tasks. Just use the approach that works for you. Think of it as practice before professional work. The main goal here is to do the analysis, in a limited amount of time, not what the codes will look like. However, if you keep your code clean according to the principles you learned in the material on "Clean and Reproducible Code Writing" you can expect to get some bonus points from me to increase your overall score above the 70pct limit. The same applies to the BONUS TASK. It is scored completely outside the 70 point limit. There is no answer key to the bonus task either. If I like your answer (and your thinking), I will give you some bonus points. This assignment is designed to mimic real work situations. Usually you won't get a set of tasks, but rather a puzzle to solve using data and your coding skills. Good luck!

REMEMBER: Communication with other students and using AI tools like ChatGPT, Copilot, Bing, etc. is **strictly forbidden**.

Once you have solved the tasks, please submit your answers via the form below:

https://forms.gle/HaNuRezwQgQgdtKf7

Your tasks:

Welcome to the Galactic Travel Agency, an interstellar hub for voyagers across the universe. In this scenario, you are a data analyst tasked with unravelling the complexities of cosmic tourism. Our vast database contains intricate details about numerous galactic journeys, encompassing a diverse clientele from various species and planets. Your mission involves sifting through this cosmic data to identify trends, preferences, and the economics of space travel. Delve into the mysteries of the galaxy and uncover what drives the interstellar traveller. Are luxury trips to Andromeda more popular among

Martians or Jovians? Does the duration of a voyage affect its price? Use your R skills to explore these astronomical queries and more, as you chart a course through this star-studded dataset.

To access the dataset, please use the following code:

read the csv file galacticTravel.csv (be careful of about the separator and decimal mark!)

use filtering on the rows, depending on your student ID number (if you're an Erasmus student, use only the numbers in your ID)

```
data <- read.csv … # READ DATA HERE

# IMPORTANT!! Remember about separator and decimal!!!

id <- 123456 # YOUR ID HERE

set.seed(id)

myData <- as.data.frame(data[sample(1:10000,500,replace=FALSE),])
```

Description of the dataset:

TravelDate	Day of the travel
Destination	Destination of the travel, which includes: the galactic destination (Marus, Venus, Jupiter, Saturn, Andromeda) and the information whether travel ends on the actual planet (P) or the galactic station in the orbit (S)
Species	Species of the traveller
TravelClass	Class of the travel (economy, business, luxury)
Duration	Duration of the travel in full Standard Galactic Days
PromoTokens	Amount of promo tokens applied for this travel (discount points from the loyalty programme)
Price	Final price paid by the traveller

Tasks:

1. Read the description of the dataset and compare it with the result of the str() function. What types/classes are assigned to the variables (currently) and how they should be represented in R to ensure the most efficient and convenient calculations (eg. factor, numeric, character…)?

	Type (currently)	Target type/class in R
TravelTime		
Species		
TravelClass		
Duration		
Price		

2. Rename variable 'PromoTokens' to 'PromotionTokens'. Correct the error created at the data creation stage. Do not create a new variable, just rename an existing one.

3. Examine the contents of the ‘Destination’ variable. What type of information does it store? Separate two pieces of information from this variable into the variables ‘DestinationPlace’ and ‘StationOrPlanet’.
4. Transform the ‘TravelDate’ variable to the appropriate type. Make sure your changes affect the data.frame that you are operating on.
5. Among the travels, which received at least 8 promotion tokens, assess how individuals of different Species undertook them (TIP – remember how to take a look inside of a factor variable).
6. Add a new variable to the dataset, which will store the price of a trip calculated per one Standard Galactic Day (SGD). Name it “PricePerDay”. TIP – divide (and conquer ☺).
7. Calculate the median price of travel to Venus in Luxury class.
8. Create a plot which will store TWO histograms side by side, which will show the distribution of travel time towards Mars and Venus. Change the color of bars for the first plot to “red”, and color of bars in the second plot to “blue”.
9. Build a linear model “myModel” (function lm()) that explains the price of the travel as a function of the trip destination (both variables if you did the task 3 successfully), travel class, travel duration, used promotional tokens and Species that undertook the travel. Extract the coefficient for trip duration and print it out.
10. Create a group comparison in the form of a summary table (which should look like the one below):

Species	TravelClass	averagePrice	maxPromoTokens
Andromedan	Business	X	X
Andromedan	Economy	X	X
Andromedan	Luxury	X	X
Human	Business	X	X
Human	Economy	X	X
Human	Luxury	X	X
Jovian	Business	X	X
….	….	…	…
….	….	…	…

It should compare the average price of travel and the maximum promotional token used across different travel classes and species of the travellers. TIP: ordering (and number) of the rows may be different than in the example.

11. Create a new variable MySpeciesAverage which will contain the average price of travel for given Species. Values of this variable should be the same across the observations within the same species group but should differ if we compare rows of different species. Bonus points if you create this variable using one line of code.

BONUS TASK, BONUS POINTS: Using your expert skills, the power of the galactic flow and the R code, prepare a more optimal pricing strategy for intergalactic travels offered by the Galactic Travel Agency. Explain your decision and way of thinking (is it data-driven? Or is it just your sense of intergalactic trends?). Using the results of the linear model from the previous tasks, explain how your pricing strategy may differ from the current relationships between travel characteristics and travel price. Keep your explanations short - the written part of this assignment (without code) must be no longer than 200 words.