RIntro Exam – 1st term 2025/26 – group 14:00
by dr Maria Kubara

Item	Condition	Seller Rating	Dias Listed	Final Price
Designer Bag	Like New	★★★★★	12 Days	$150
Vintage Dress	Good	★★★★★	30 Days	$150
Vintage Dress	Good	★★★★	3 Days	$45
Sneakers	Fair	★★★★	5 Days	$20

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. Any suspicion of such communication or AI usage will result in immediate invalidation of the exam and may lead to formal disciplinary proceedings, in line with university regulations.

Once you have solved the tasks, please submit your answers via the form below:
https://forms.gle/nkP2Q8RsyZ1BY9TW9
This time, you **do not need to worry about white spaces** or code formatting.
Just make sure to paste only the code relevant to a given task.
Concise, clean, and working solutions will be scored higher than long or “spaghetti-like” code. By “spaghetti code” we mean code that is unnecessarily long, tangled, hard to read, and difficult to follow, even if it technically works. Keep that in mind.

Your tasks:
An online marketplace connects buyers and sellers of second-hand clothing. Each row is one completed sale. Your job is to analyse a sample of transactions and extract simple insights about price formation, seller behaviour, and marketplace dynamics.

To access the dataset, please use the following code as your template:

```
# read the data file secondhand_fashion.csv (be careful 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 DATA HERE
  id <- … # YOUR ID HERE
  set.seed(id)
  myData <-as.data.frame(data[sample(1:10000,500,replace=FALSE),])
```

After making appropriate changes to the function reading the data and your ID number paste the code above (and only this part of the code!) in the first slot in the Google Form.

Description of the dataset:

- *SaleDate* – date of sale.
- *ItemDetails* – combined info about item category and condition (e.g., Shoes_Good, Dress_New).
- *BrandTier* – one of the three tiers: Low / Mid / Premium.
- *SellerRating* – seller rating (in range 1–5).
- *DaysListed* – number of days the item was listed before sale.
- *ShippingMethod* – three categories: Economy / Standard / Express.
- *DiscountApplied* – whether discount was applied (Yes/No).
- *FinalPrice* – final sale price in USD.

Tasks:

1. Read the description of the dataset and compare it with the current datatypes in the read database. What types/classes are assigned to the variables (currently) and how they should be represented in R (sometimes after additional transformations) to ensure the most efficient and convenient calculations (eg. factor, numeric, character…)?

	Type (currently)	Target type/class in R
SaleDate		
BrandTier		
SellerRating		
DiscountApplied		
FinalPrice		

2. Rename the variable ItemDetails to ListingInfo. Modify the dataset directly without creating a new variable.

3. Examine the contents of the ListingInfo variable. Separate its information into two new variables: Category (e.g., Shoes, Dress, etc.), Condition (e.g., Good, New, etc.).

4. Transform the SaleDate variable into a standard date format (YYYY-MM-DD). Ensure your changes affect the dataset directly.

5. Clean the FinalPrice variable so that all mixed data types are removed. The target is for the variable to contain only numeric values, ready for further calculations. Do not create a new variable — clean the existing one by removing the unit.

6. For transactions where sellers were rated with more than 4.6, calculate the average FinalPrice.

7. Add a new variable to the dataset named PricePerDayListed, which divides the final price by the number of days the offer was available in the system (use appropriate calculations).

8. Find the maximum FinalPrice among items that are new and marked as premium in the brand tier.

9. Create three histograms displayed side by side using base R plotting to compare the distribution of FinalPrice for different values of BrandTier. Adjust the plotting window so that the histograms appear in a single row. The first histogram should show products in the Low tier, the second in the Mid tier, and the third in the Premium tier, with titles "Low", "Mid", and "Premium", respectively. Colour the histogram bars gray, label the x-axis as "Final price", and leave the y-axis label empty. After creating the plots, restore the plotting window to its default settings.

10. Build a linear model myModel explaining FinalPrice using predictors: BrandTier, Category, Condition, DaysListed, SellerRating, DiscountApplied, ShippingMethod. Extract the coefficient for DiscountApplied and explain its meaning (in plain language).

11. Create a summary table comparing: AveragePrice (mean FinalPrice) and MaxDaysListed (max DaysListed) across groups of Category and Condition. The table should resemble the following structure:

Category	Condition	AveragePrice	MaxDaysListed
Accessories	Fair	X	X
Accessories	Good	X	X
Accessories	VeryGood	X	X
Accessories	New	X	X
Dress	Fair	X	X
….	….	…	…

TIP: ordering (and number) of the rows may be different than in the example.

12. In the dataset myData, introduce missing values in the variable FinalPrice by setting every 32nd observation to NA using indexing. Next, impute these missing values by replacing them with the median of FinalPrice. When computing the median, remember to handle missing values properly.

BONUS TASK, BONUS POINTS:

Using your analyst intuition and your results, suggest one change to platform rules/design that could increase seller revenue without making buyers worse off. Explain whether your recommendation is mainly based on evidence from your analysis or your intuition. Keep your explanation concise — the written part of this assignment (without code) must not exceed 200 words.