LAB 1: R basics & refresher

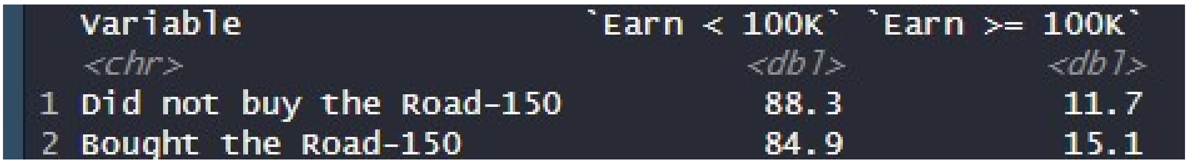
Data comes from AdventureWorks, a sample of a database made by Microsoft used for training to design SQL server databases, using data from the Adventure Works brand.

There are 10 tables in the database:

- Calendar: with a range of dates
- Customers: customer data, with their name, income, number of children, education level…
- Products
    - Main table: information about products sold by AdventureWorks, name of product, cost, color, price…
    - Subcategories: categories of products
    - Categories: categories of products
- Returns: returned products
- Sales (2015, 2016 and 2017): sales with the information about the product, customer who bought it, how many, date and where it was bought
- Territories: information about geographical locations (region, country, continent)

Exercise

1. Import all the .csv tables and have a look at the data. Quickly describe each table. How many rows and columns? What is the table for?
2. There are three sales tables. Bind them all into one table. How many orders have been made over those three years?
3. Merge the three products tables: products, products_categories and products_subcategories. What key(s) did you use to join the tables?
4. Starting from your new sales table, join it to your new products table, with the customers table, and the territories table. How many rows are there? Why do you think the tables were split up in the first place?
5. Starting from the customers table, left join it to the sales tables, the territories table, and the products table. How many rows are there? Why is it different from the number of rows you had in the table from the previous question? What would you change or do to get the same number of rows?
6. How many customers earn over $100,000? On which table should you check this?
7. Is there missing data in the Customers table? How many rows have missing data? In which variable(s)?
8. How many orders have been made from people with income over $100,000?
9. What is the average price of products bought by people with an income over $100,000? Under $100,000?
    a. Calculate the 95% confidence interval of these averages
    b. Also, do a t-test of difference in means (:
10. Do a histogram of the distribution of prices of the models sold by Adventure Works
11. What is the most expensive bike sold by the company? Among all customers, how many bought it? Give the number and the percentage.
12. Are there more high-earners who bought the most expensive bike, compared to lower-earners? Do a two-way table, using R (not Excel!), in this format:



```
  Variable                  `Earn < 100K` `Earn >= 100K`
  <chr>                             <dbl>          <dbl>
1 Did not buy the Road-150           88.3           11.7
2 Bought the Road-150                84.9           15.1
```

The earnings variable should be in column, the binary variable indicating if the customer bought the most expensive bike in row, and row percentages* should be shown.

*Row percentages means that the percentages in a row should sum to 100.

13. Export this table to a .csv (or .xlsx) file.
14. Do a chi-squared test of independence to check if there is an association between high earnings and buying the most expensive bike.

Bonus (if you have time)

Calculate total profit per customer. Use the GLM of your choice to regress total profit on the variables you deem relevant. Quickly interpret your results.