Week 2 | Examination assignments

Data acquisition, import, and manipulation (wrangling)

Erik Vanhatalo

2025-03-16

Contents

Regarding the Assignments

This document outlines the **individual** examination assignments for this week. We strongly recommend that you work through the training assignments for this week before progressing to the examination assignments.

Prepare and document solution before the examination session

You need to prepare your solutions before the examination session! At the examination session you will need to have your imported datafiles and prepared solutions easily available so that you will be able to perform the tasks you are given at the session.

Quarto - recommended way of documenting your solutions

There are some options for you to prepare, document and save your solutions. You can do it in a standard R script (.R file). which is easy and straightforward. However, we strongly recommend that you use Quarto because it makes it possible for you to combine executable code with your own notes. You can also create *MS Word*, *pdf* or *html* documents with your solutions that you can use as you prepare. During Week 1 there will be an introductory lecture featuring Quarto. Quarto is integrated in RStudio. **Whatever choice you make it is important that you are able to run your prepared solutions fast and easy at the examination sessions.**

1. Examination assignment 1

You will be working with the MS Excel dataset: **"Online Retail II".** This is a big dataset (with many rows) containing all transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company's products are mainly unique all-occasion gift ware.

1.1 List of tasks

You should complete the following tasks or answer the following questions:

1. Import the dataset into Microsoft Power BI.
2. Import the dataset into R and make sure that all variables/features have the correct class.
3. Change the names of the two variables/features:

- Change the '*StockCode*' variable to 'Product_number', which is easier to understand.
- Change the name of '*Customer ID*' to '*Customer_ID*', which is good from an R-perspective eliminating empty space in variable name.

4. What is the dimensionality of the dataset. How many observations and variables/features does it have?
5. Answer these questions:
    - How many **unique customers** are in the dataset?
    - How many **unique products** have been sold during these years?
6. To how many countries and to which countries have orders been shipped during these years?
7. Diagnose missing data.
    - Are there missing data in the dataset?
    - If so, which variables/features have missing data?
8. Construct and name a new dataset where all rows that contain missing values are removed.
    - How many rows(observations) were removed?
9. For this new dataset without missing values, construct a box plot with an illustration of the "*Price*" of the orders in the dataset.
10. For this new dataset without missing values, are there any outliers for the "*Price*" variable/feature?
    - If so, illustrate and discuss the outliers.
11. For this new dataset without missing values, bin the "*Price*" variable/feature into the following classes:
    - "small order", "intermediate order", "large order", and
    - then **add these bins/classes** to the dataset without missing values.
12. Given your created bins, how many "small", "intermediate", and "large" orders are there?

2. Examination assignment 2

You will be working with the .csv file **"winequality-red.csv".** You should read up more about the dataset in the associated information file from UCI machine learning repository.

2.1 List of tasks

You should complete the following tasks or answer the following questions:

1. Import the dataset into Microsoft Power BI.
2. Import the dataset into R. Make sure that all variables/features have the correct class.
   - Fix the comma delimiter issue!
3. Change the variable names into something that is easy to understand but do not have empty spaces in the variable names. This will make "life" in R easier.
4. Diagnose outliers in the variables/features. Which three (3) variables have the highest ratio of outliers?
5. Find one way (you should be creative) to illustrate the correlation between the quality score and all other eleven variables/features.
   - Your solution needs to show how the quality score of the wines correlate with the measured variables.
   - Which feature is **most strongly negatively** correlated with the quality score of the red wines?
   - Which feature is **most strongly positively** correlated with the quality score of the red wines?
6. Create a subset **dataset including only wines** with quality scores six (6) or higher.
   - How many rows/observations or wines fulfill this requirement?

3. The new assignments at the examination session?

If you have solved the above tasks for the two assignments, documented your solutions, and are able to run the code fast in the classroom you should be well-prepared to solve the additional assignment(s) that you will get at the examination session. In preparation for the examination assignments it is also important that you and have worked through the training assignments. If you have done all this, **do not worry!**