Київський національний університет імені Тараса Шевченка

Факультет комп’ютерних наук та кібернетики

Лабораторна робота

з курсу

«Аналіз даних»

на тему

**«Кореляційний аналіз даних.»**

Виконав:

студент групи ПІ-23

факультету комп’ютерних наук

та кібернетики

Євчик Олексій Юрійович

Київ 2023

***Зміст***

[Опис вхідної інформації](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.gjdgxs)

[Постановка задачі](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.30j0zll)

[Перевірка змінних на нормальний розподіл](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.1fob9te)

[Аналіз істотності парних статистичних зв'язків](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.3znysh7)

[Аналіз істотності множинних статистичних зв'язків](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.tyjcwt)

[Графічне представлення](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.3dy6vkm)

[Підсумок](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.1t3h5sf)

[Список використаних джерел](https://docs.google.com/document/d/1k_B3p8r30TUZ7FaWy8ak7gWKzdMWpkN2/edit#heading=h.4d34og8)

***Опис вхідної інформації***

# Я досліджував набір даних «Fastfood Nutrition», опублікований на сайті kaggle.com( <https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition?resource=download> ). Цей датасет надає повну інформацію про поживну цінність різноманітних продуктів швидкого харчування від популярних мереж швидкої їжі. Фаст-фуд відомий своєю зручністю та доступністю, але він також сумно відомий своєю високою калорійністю, високим вмістом жиру та цукру. Цей набір даних має кілька скалярних змінних: calories (калорії), cal\_fat (жири), total\_fat (всі жири), sat\_fat(насичені жири), trans\_fat(транс-жири), cholesterol (холестерин) , sodium (натрій), total\_carb (вуглеводи).

# Я буду використовувати три змінні: calories (калорії), total\_carb (вуглеводи), cholesterol (холестерин). Всі три змінні є кількісні скалярні. Серед рядків датасету наведено різні продукти( включно з підвидами, наприклад: сир моцарела й салат цезар). Всього він нараховує 515 різних продуктів.

# *Постановка задачі*

Нам потрібно провести кореляційний аналіз набору даних

Для кожної зі змінних **calories (**калорії**), total\_carb (**вуглеводи**), cholesterol (**холестерин**)**:

На основі результатів попередньої обробки обраного набору даних визначитися, які характеристики статистичного зв'язку потрібно використати при подальшому їх кореляційному аналізі,

Провести аналіз істотності парних статистичних зв'язків для усіх пар скалярних змінних, навівши для них:

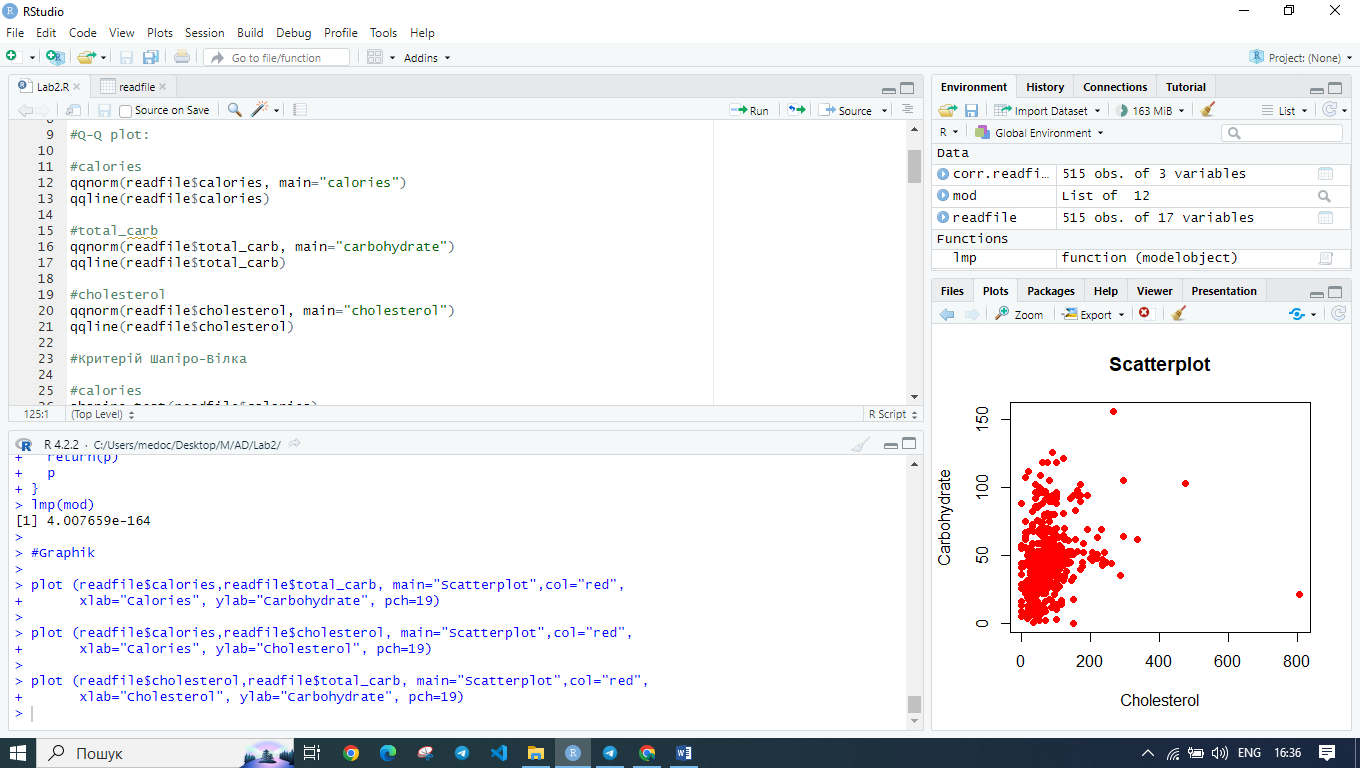
* вибіркове значення відповідної парної характеристики статистичного зв'язку,
* максимальний рівень значущості при якому відповідний парний статистичний зв'язок не є значимим,
* впорядковану послідовність усіх пар скалярних змінних у порядку спадання істотності статистичного зв'язку між ними,
* сформулювати висновки по кореляційному аналізу парних статистичних зв'язків для обраного набору скалярних змінних

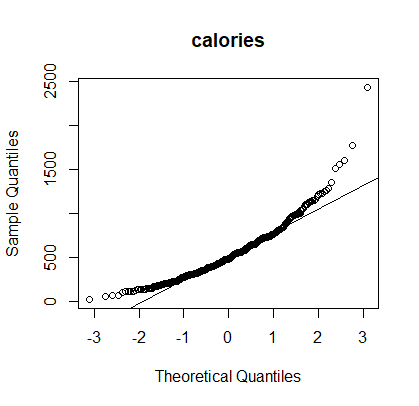
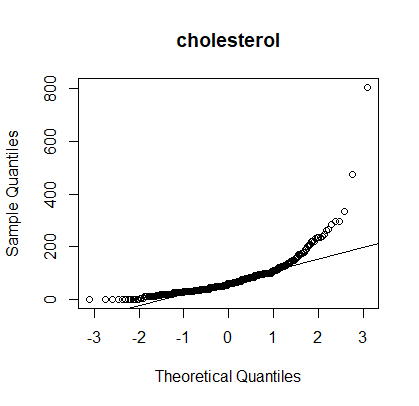
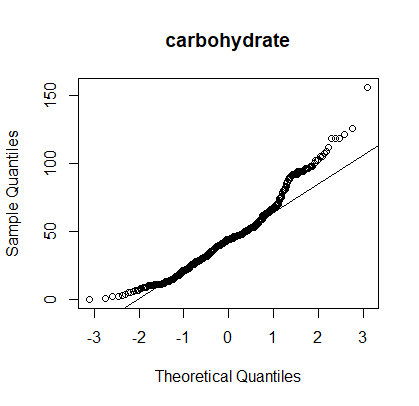
Провести аналіз істотності множинних статистичних зв'язків між кожною обраною в якості залежної скалярною змінною та множиною усіх інших скалярних змінних (які виступають у ролі незалежних змінних), навівши для них усе з попереднього списку.

***Перевірка змінних на нормальний розподіл***

Для того, щоб знати яку характеристику статистичного зв’язку використовувати, нам потрібно знати чи є змінна нормально розподіленою. Існує багато способів перевірки нормальності розподілу, тут будуть використані графічний спосіб за допомогою Q-Q plot та формальні критерії Шапіро-Вілка та Андерсона-Дарлінга.

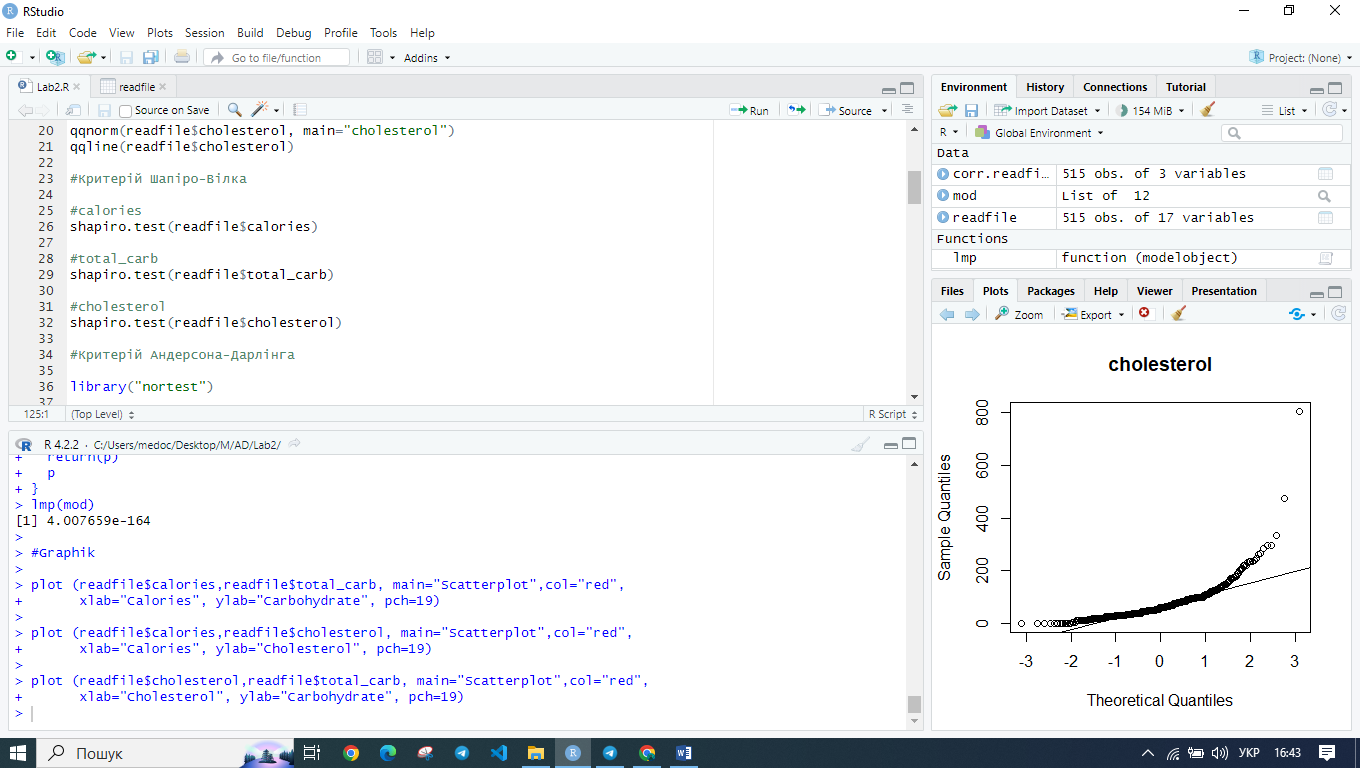
Q-Q plot дозволяє наочно оцінити рівень відхилення даних від нормального розподілу та робити певні висновки, не засновуючись лише на p-value. В даному випадку точці (х, у) на графіку відповідає один з квантилів розподілу змінної (у-координата), побудована разом з аналогічним квантилем відповідного нормального розподілу (х-координата). Пряма лінія є параметричною кривою з параметром, який є кількістю інтервалів для квантилів. В мові R графік будується за допомогою нижче зазначених команд

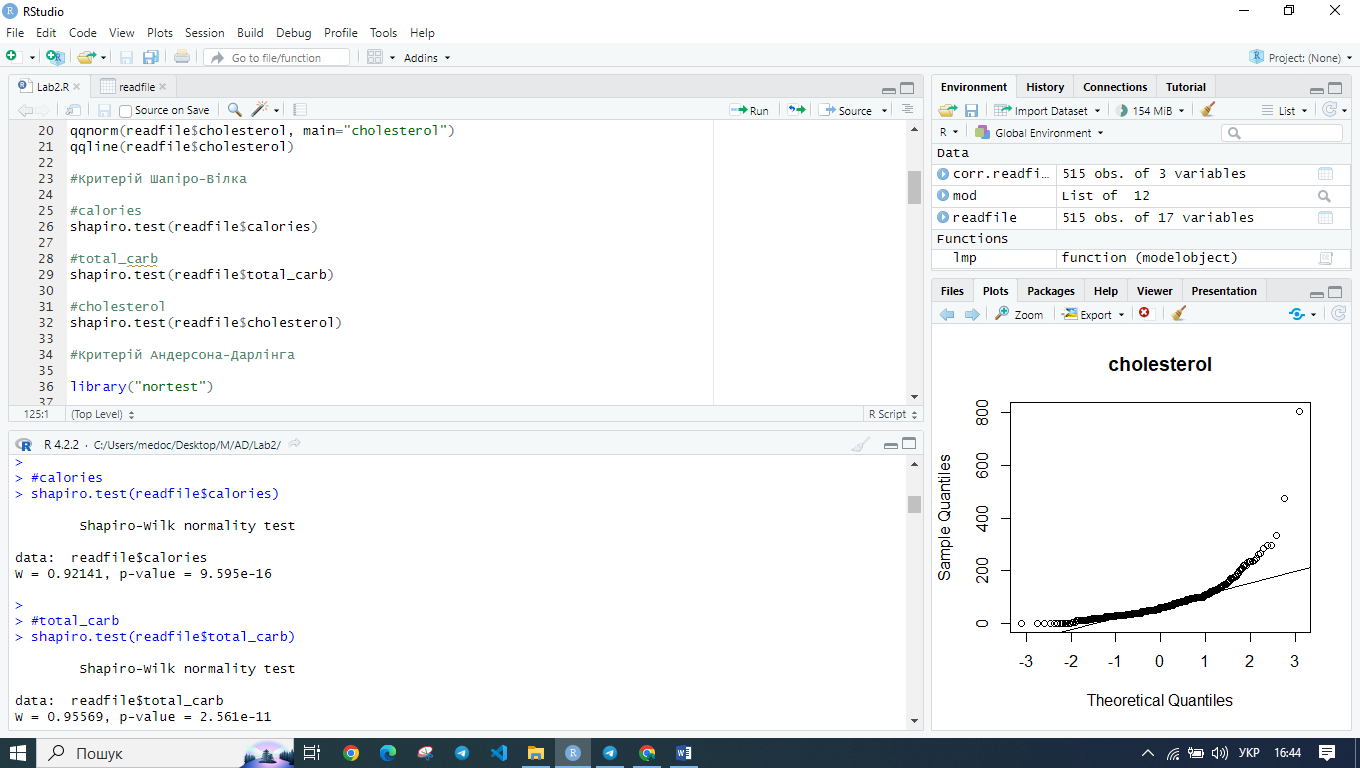


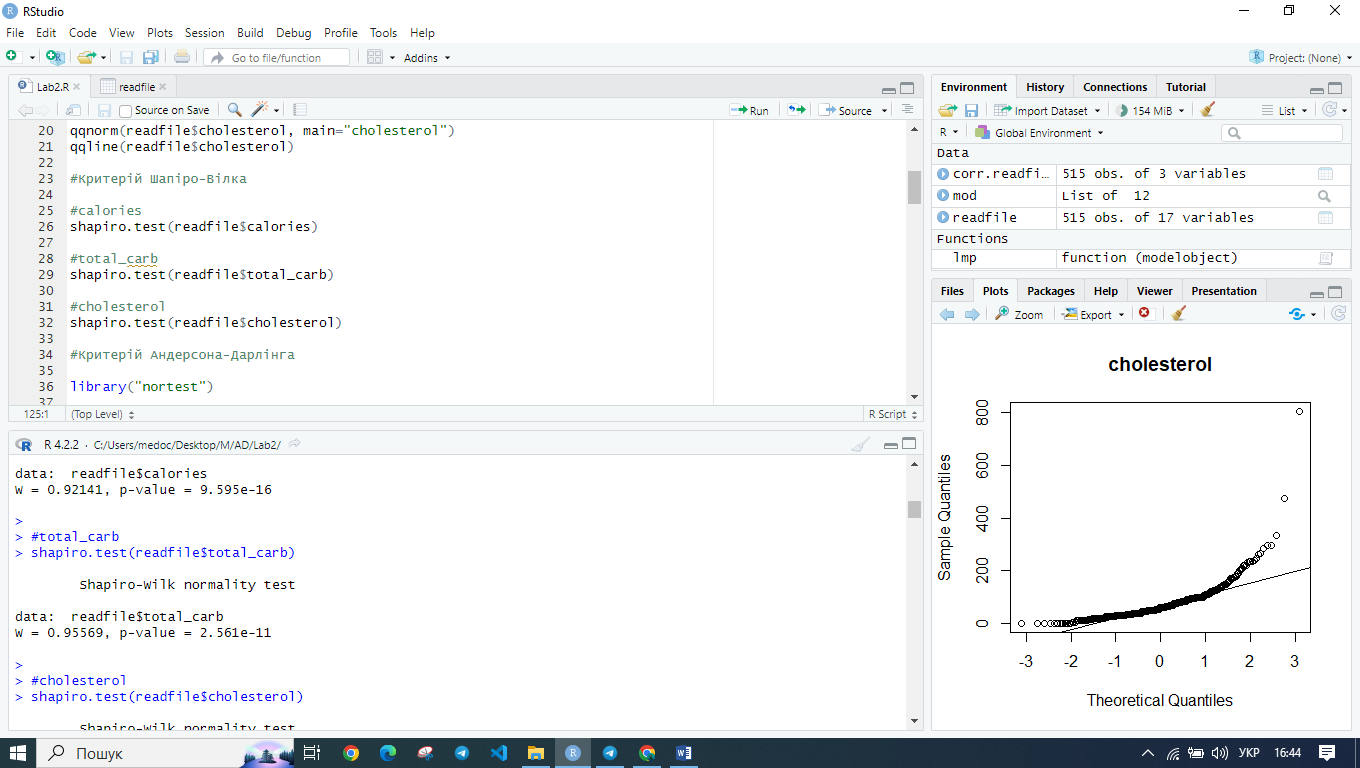
 

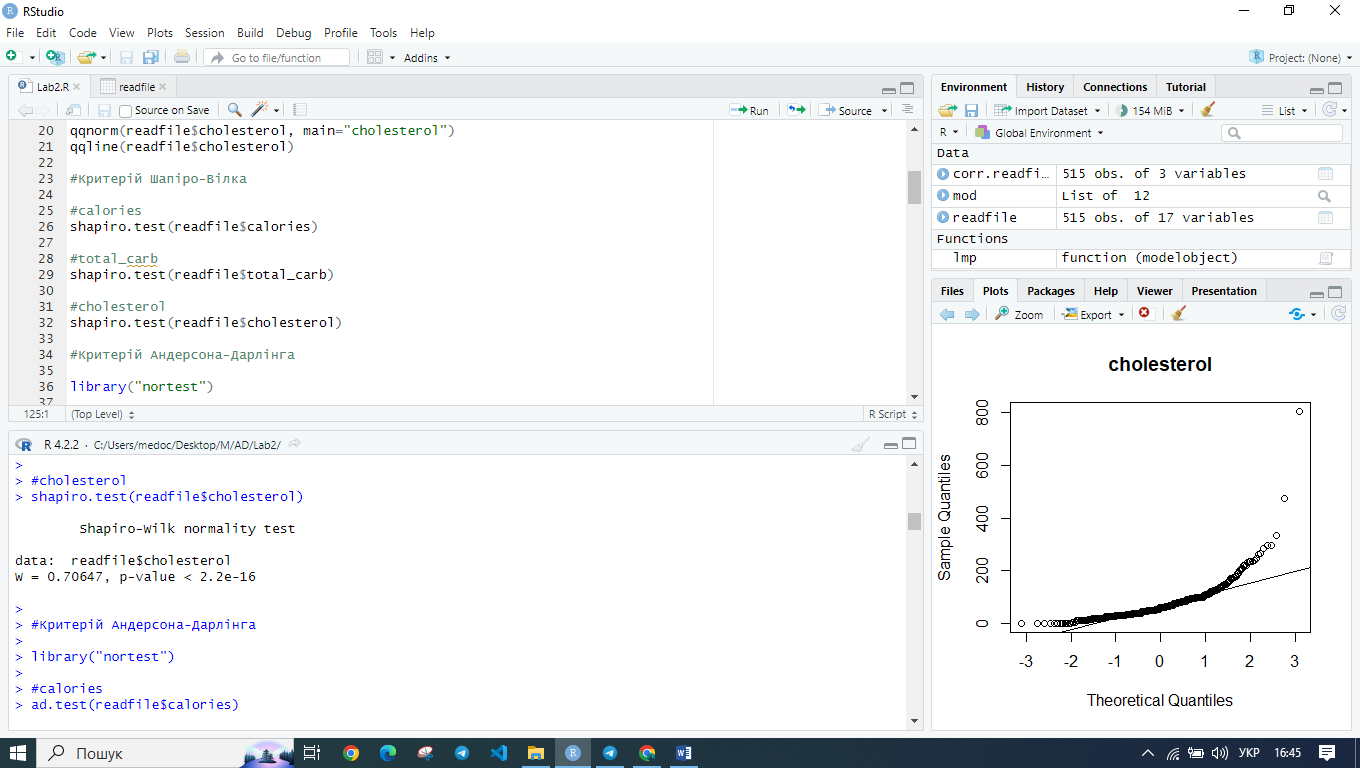
Проглянувши графіки, можна зробити припущення що змінні неє нормально розподіленими, оскільки точки на графіках розташовані не лінійно, однак це залишається лише припущенням і настав час використати критерії.

Критерій Шапіро-Вілка заключається в перевірці гіпотези Н0 про те, чи є змінна нормально розподіленою. Якщо значення p-value більше за певний рівень значущості (зазвичай 0,05), то гіпотеза приймається. Мова R дозволяє дуже зручно перевіряти ці гіпотези

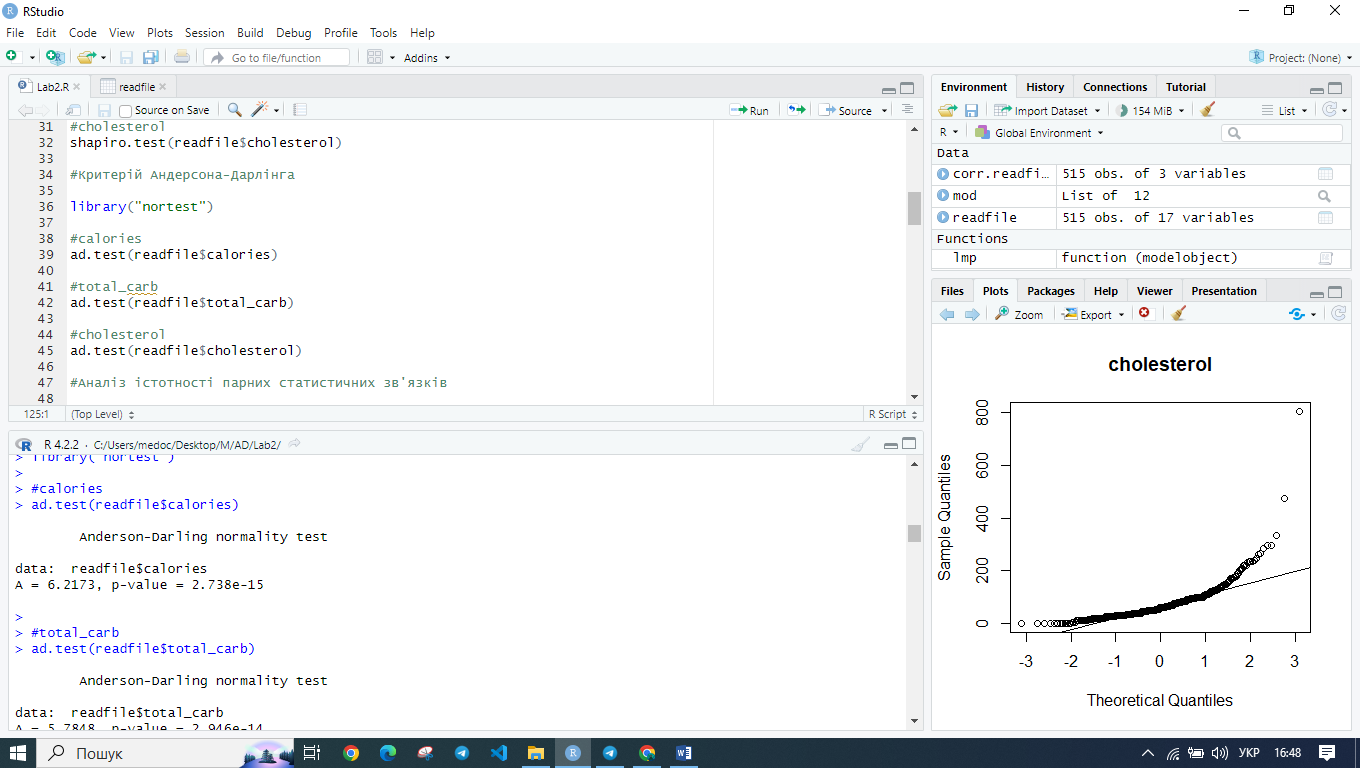


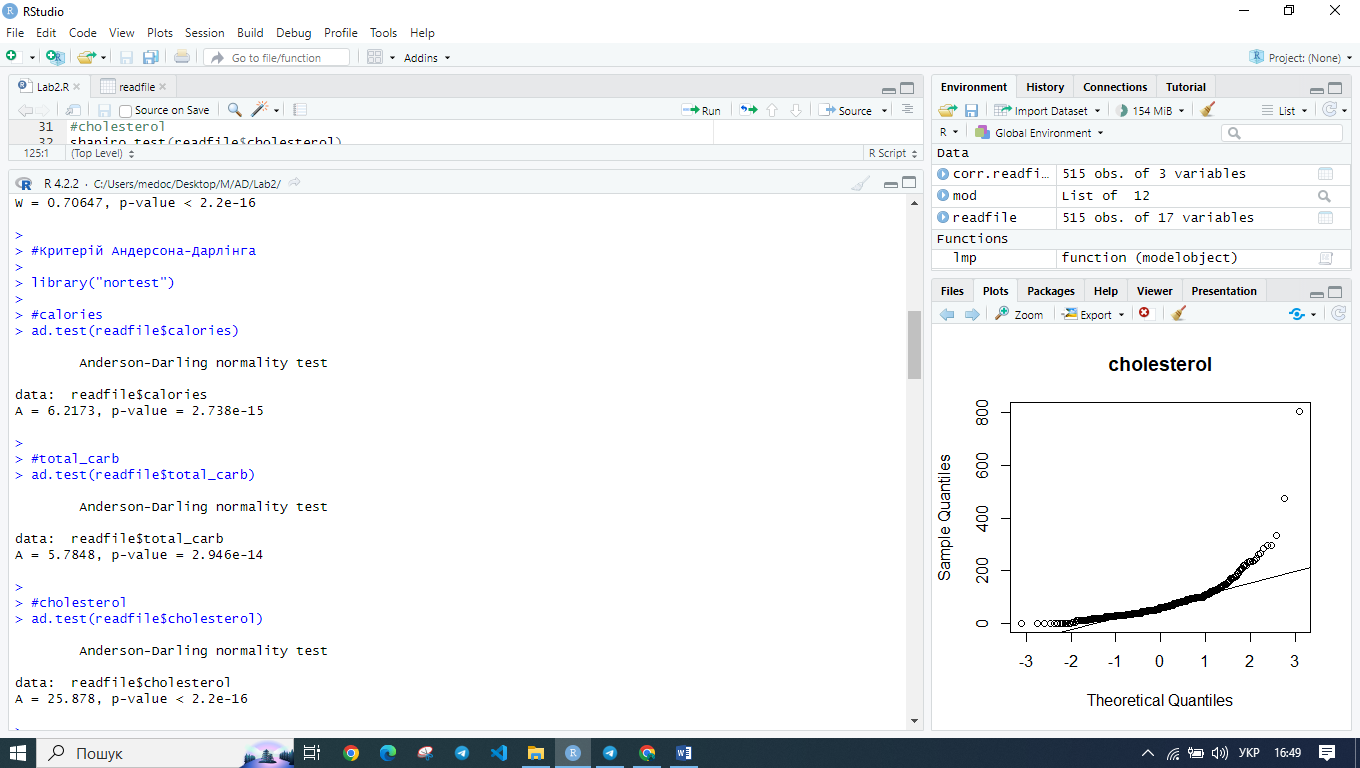






Припущення підтверджується. Проте на основі лише одного критерію Шапіро-Вілка однозначно говорити про нормальність розподілення не дуже правильно, тому перевіримо ще за допомогою критерія Андерсона-Дарлінга, який працює за аналогічною схемою. Для його використання потрібна бібліотека nortest (встановив і використав саме її).





# *Аналіз істотності парних статистичних зв'язків*

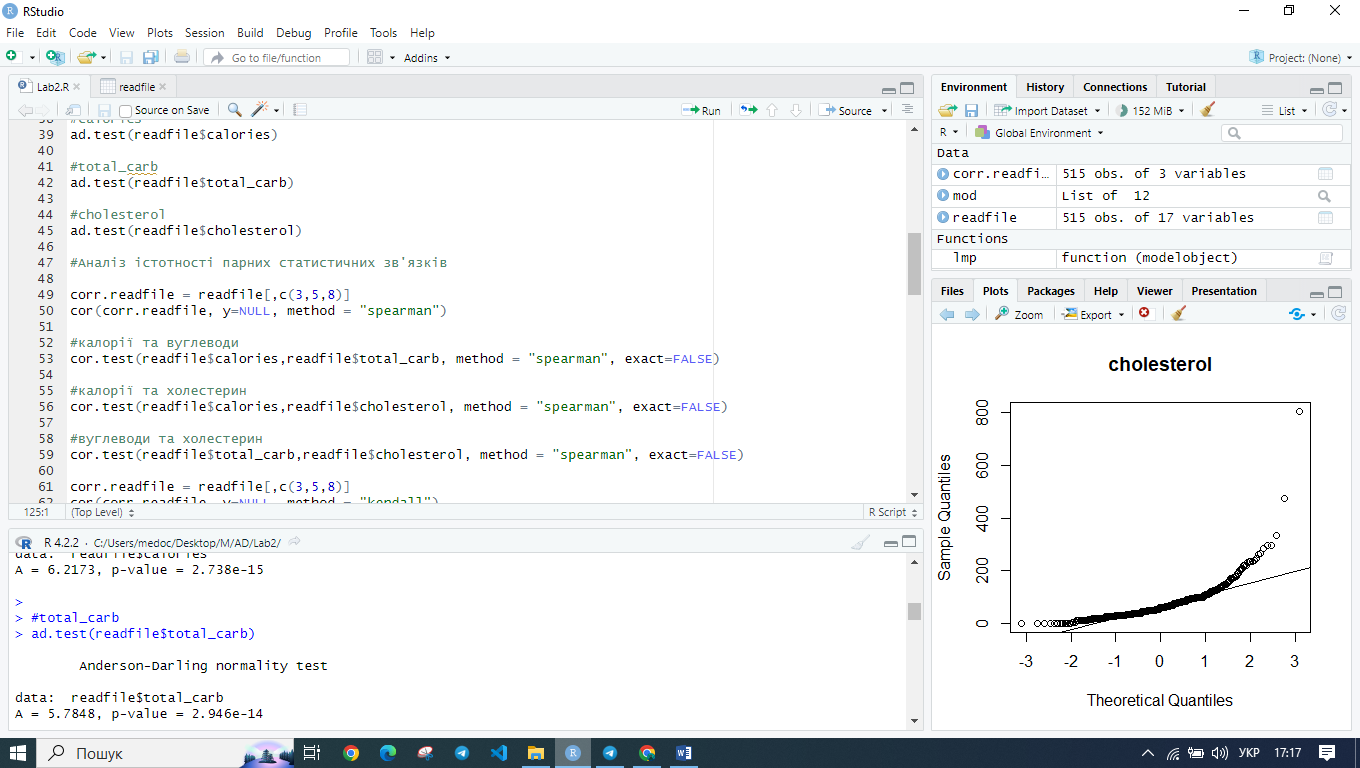
# Оскільки в нас немає нормально розподілених змінних, а отже немає пар нормально розподілених змінних, має сенс використовувати непараметричні рангові коефіцієнти Спірмена і Кендела.

# Коефіціент Спірмена обчислюється за формулою

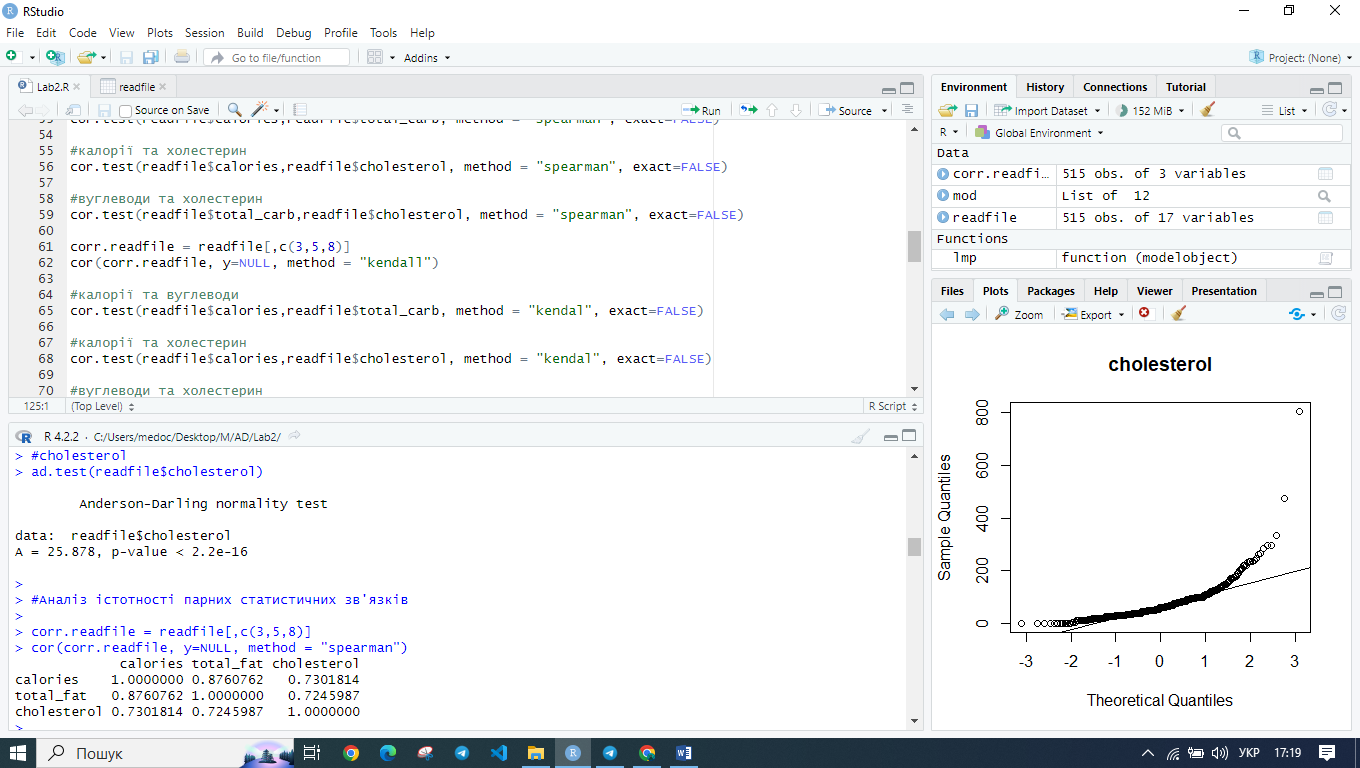
# А Кенделла за формулою:

# 

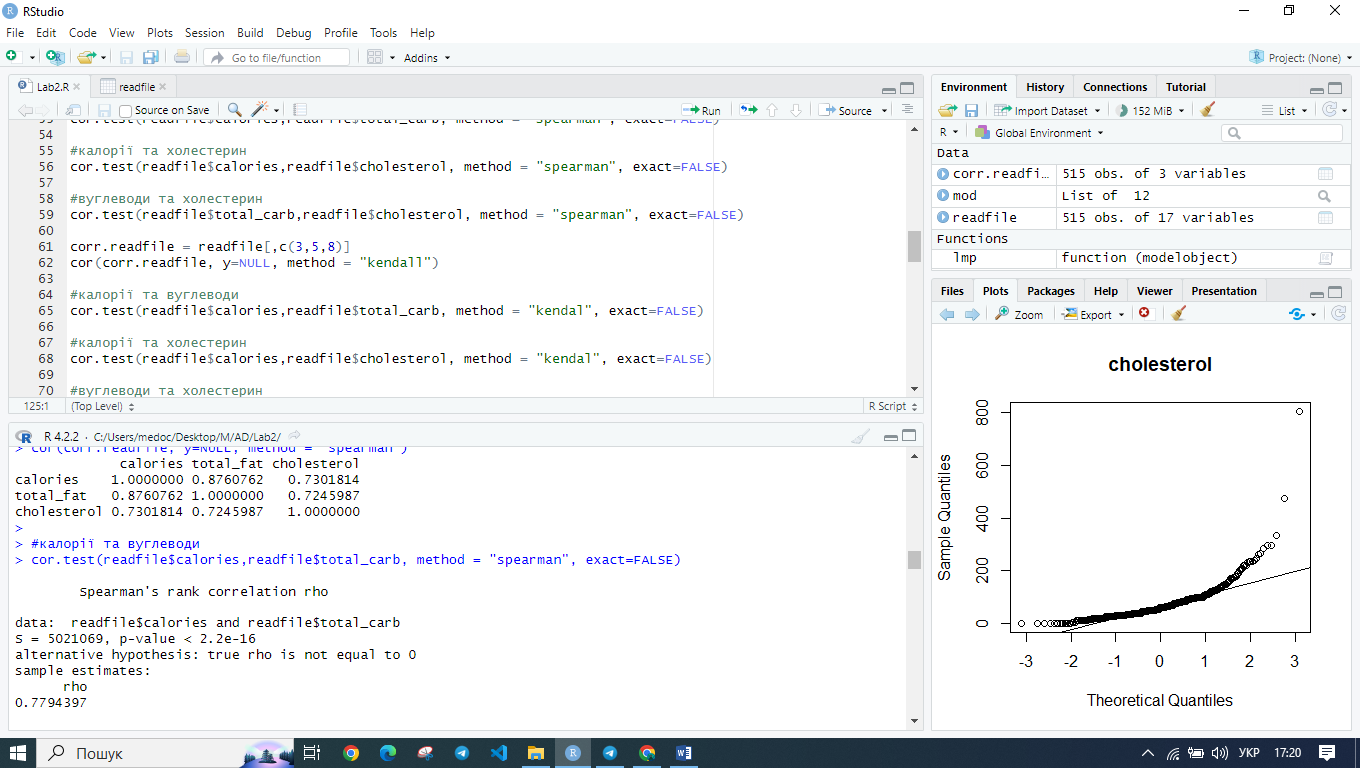
Коефіцієнти можуть набувати значень від -1 (сильна обернена кореляція) до 1 (сильна пряма кореляція). Значення, близькі до 0, вказують на відсутність кореляції.



3,5,8 – позиції вибраних зміних в файлі.



Окремо кожна пара:



# 

# 

# Тепер коефіціент Кендалла:

# 

# Й окремо для кожної пари:

# 

# 

# 

Таким чином, маємо таку послідовність пар змінних в порядку спадання істотності статистичного зв'язку між ними:

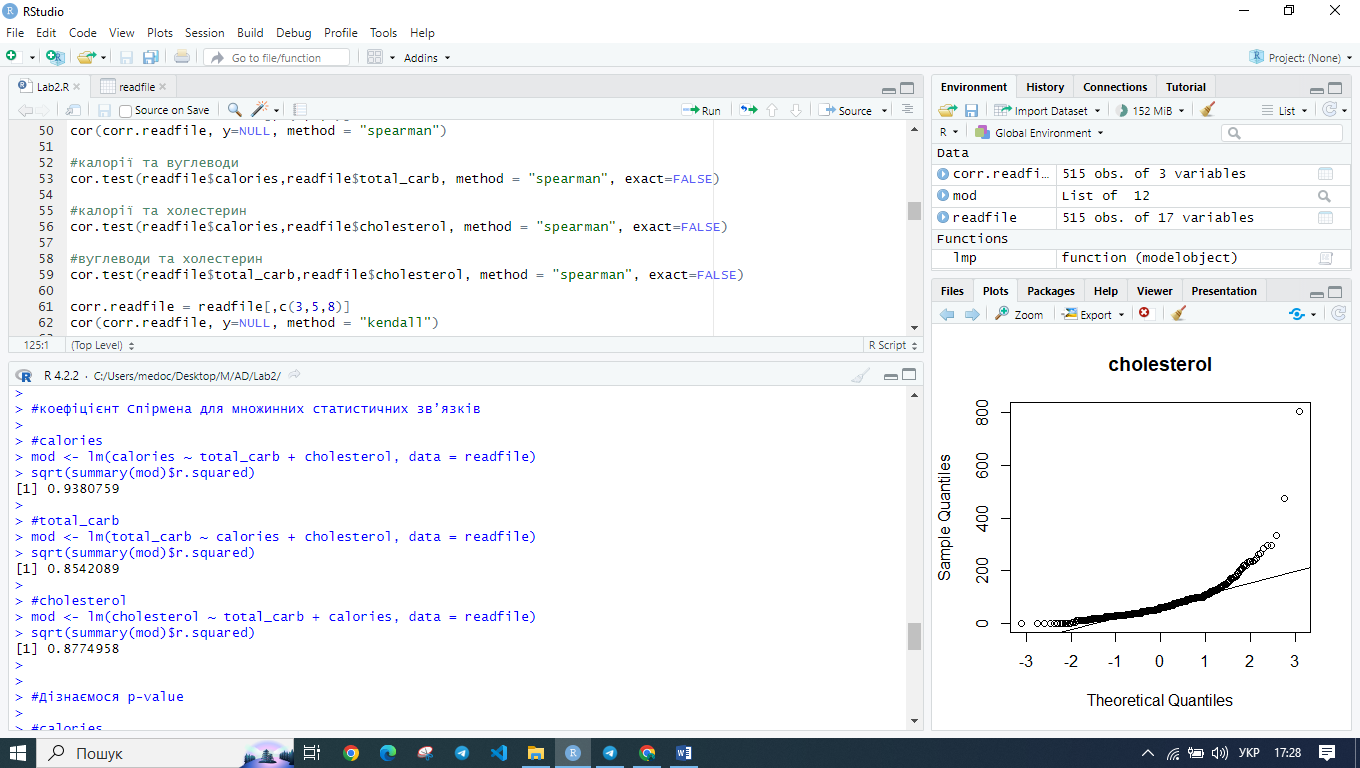
# 1 Калорії - Жири

# 2 Холестерин - Жири

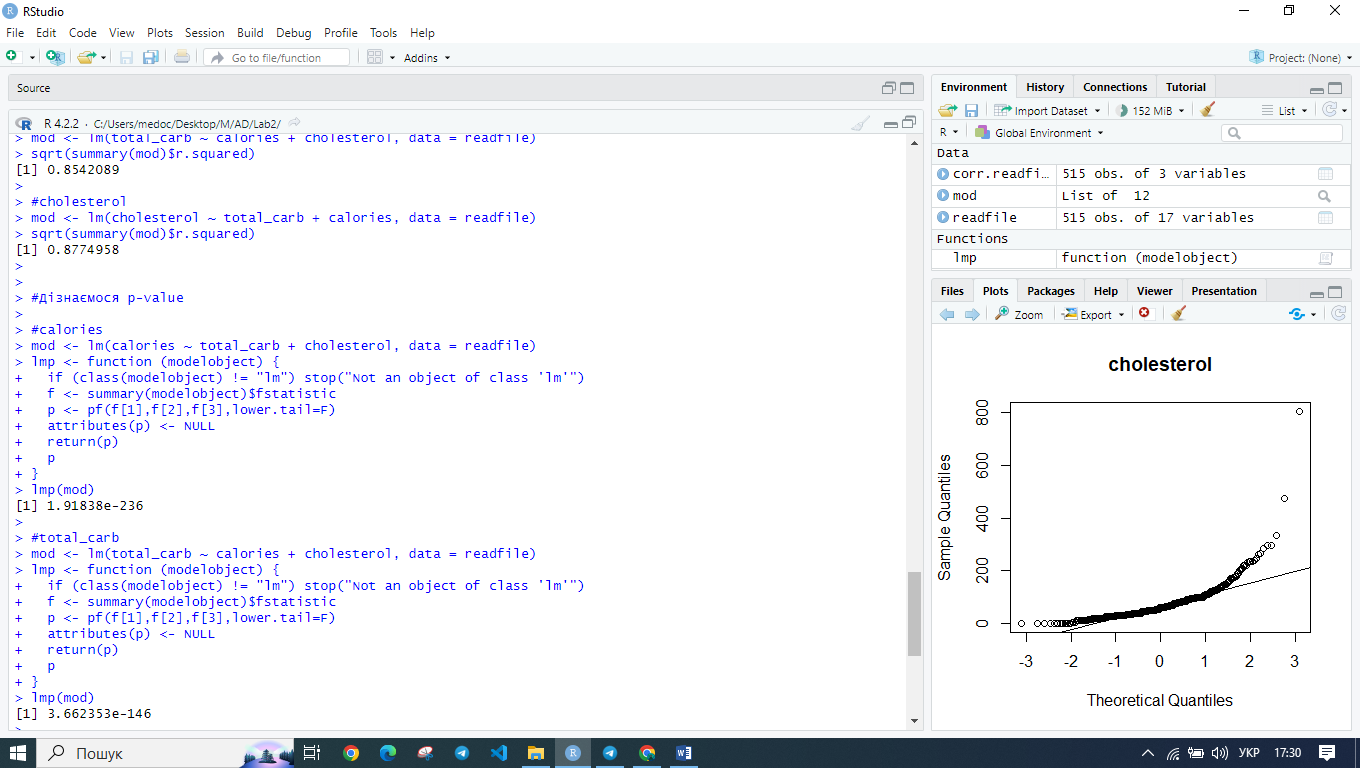
# 3 Калорії – Холестерин

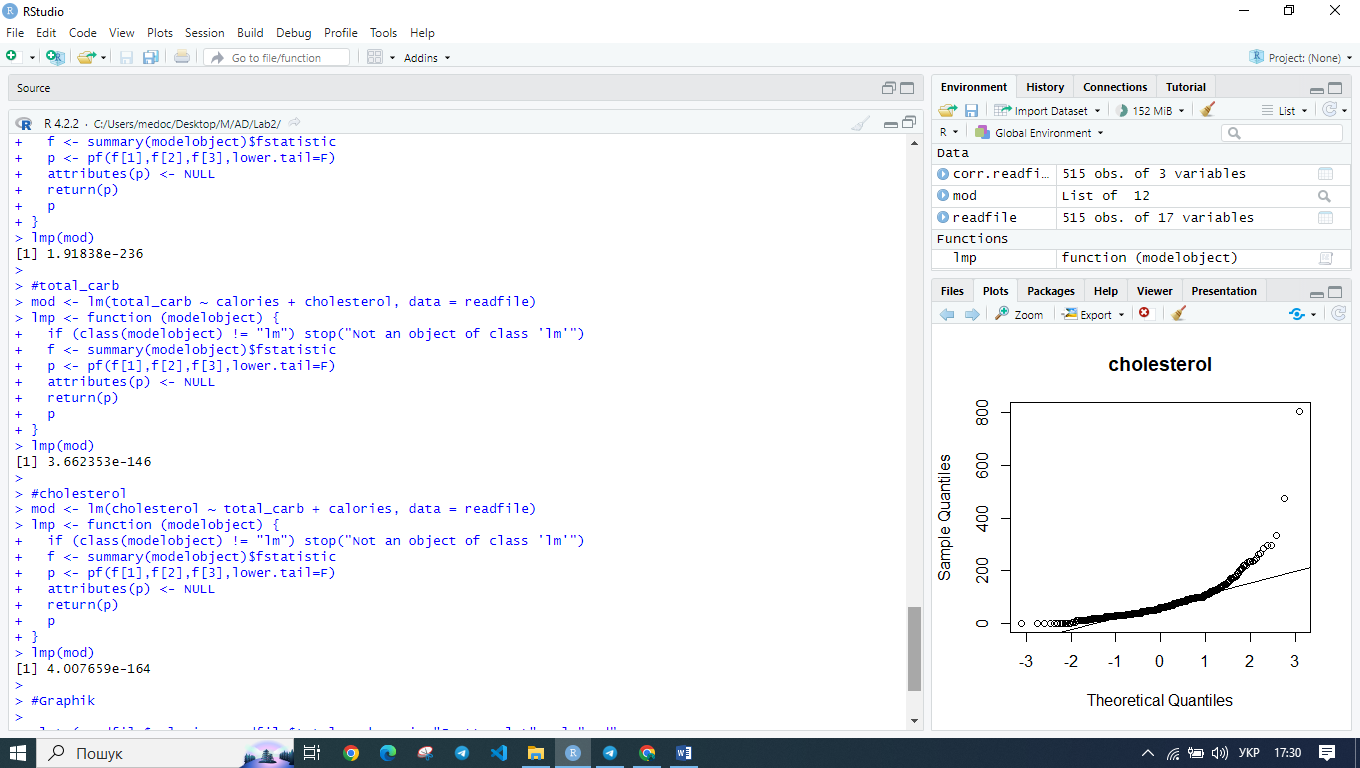
# *Аналіз істотності множинних статистичних зв'язків*

Аналогічно дізнаємося коефіцієнт Спірмена для множинних статистичних зв’язків.



Та дізнаємося p-value:





Як бачимо, кожна змінна має істотний статистичний зв’язок із множиною двох інших змінних із надзвичайно малими максимальними рівнями значущості

Таким чином, маємо таку послідовність змінних в порядку спадання істотності статистичного зв'язку між ними та множинами інших змінних:

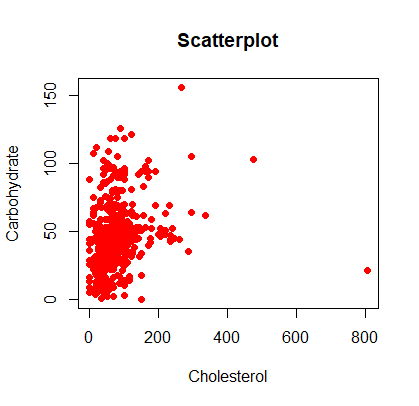
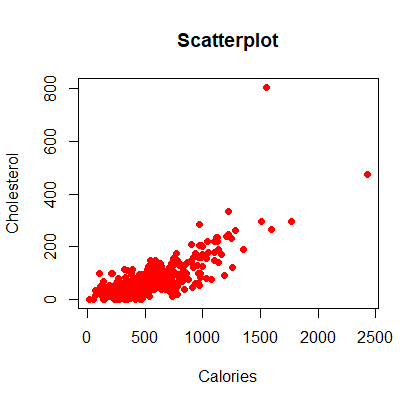
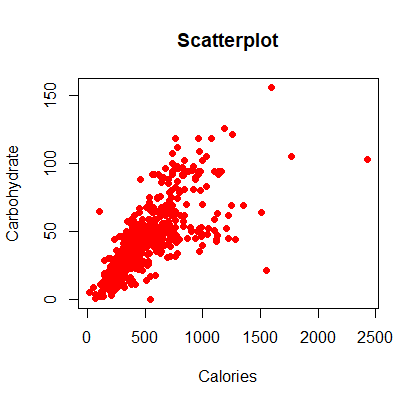
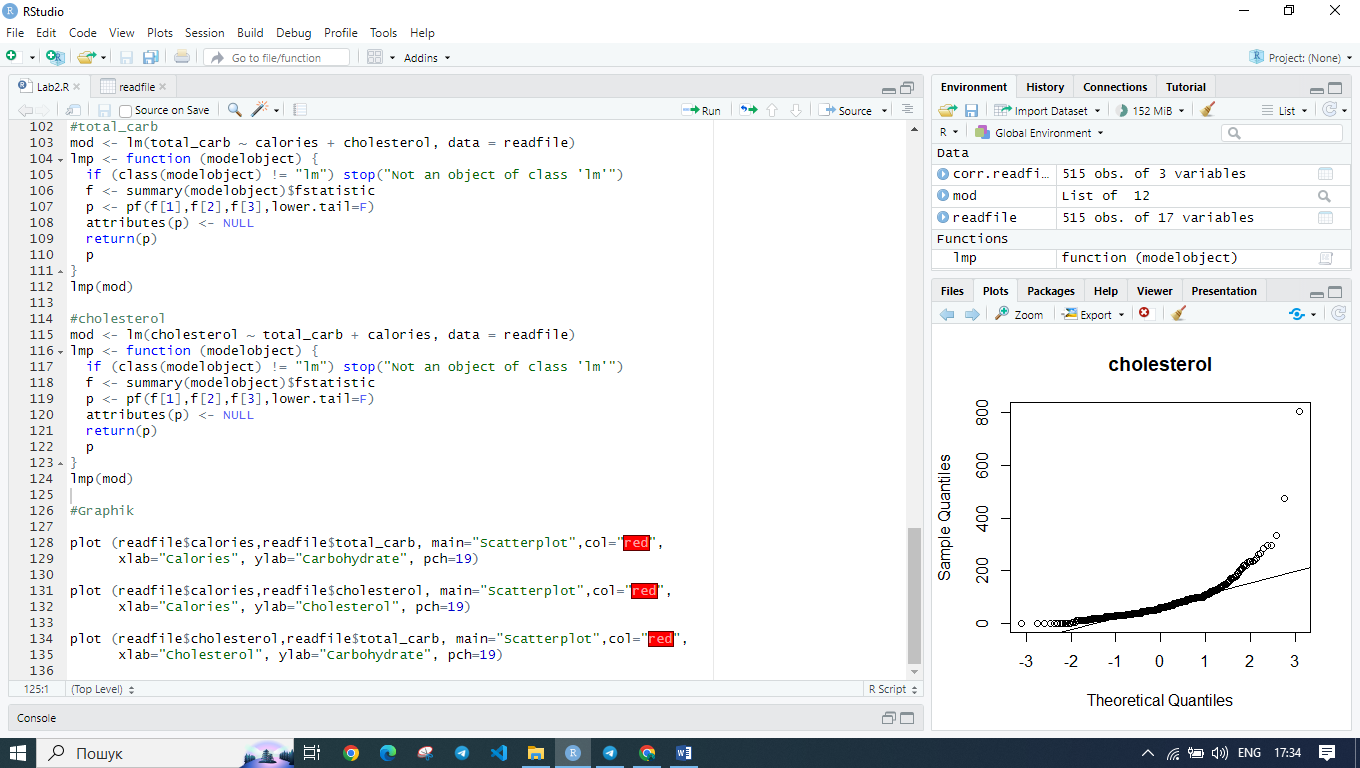
1 Калорії (calories)

2 Холестерин (cholesterol)

3 Жири (total\_fat)

***Графічне представлення***

Код мовою R:



***Підсумок***

Я провів кореляційний аналіз трьох змінних: calories(Калорії), cholesterol(Холестерин), total\_fat (Жири) із обраного мною датасету.

Перед проведенням кореляційного аналізу я перевірив розподіли змінних на нормальність за допомогою критеріїв Шапіро-Вілка та Андерсона-Дарлінга. Оскільки нормально розподілених змінних не виявилось*,* то для подальших досліджень я використовував критерії Спірмена та Кендела.

Для всіх трьох пар змінних була виявлена сильна кореляція із низьким максимальним рівнем значущості при якому відповідний парний статистичний зв'язок не є значимим (<2.2e-16). Аналогічно для всіх змінних була виявлена сильна кореляція із множинами усіх інших скалярних змінних.

Кожна з пар змінних була графічно представлена різними способами візуального представлення (Q-Q plot та діаграма розсіювання для пар), що допомогло мені наочно переконатися у правильності своїх досліджень

Дані, отримані при кореляційному аналізі, можуть бути використані в подальшому аналізі даного набору даних під час наступних лабораторних робіт.

***Список використаних джерел***

1 Лекції «Аналіз даних» Слабоспицький О. С.

2 <https://studfile.net/preview/7877820/page:19/>

3 <http://um.co.ua/8/8-7/8-7360.html>

4 <https://msn.khmnu.edu.ua>

5 <https://uk.wikipedia.org>

6 [https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition](https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition?resource=download)

7 Книга Основи Кореляційного аналізу 2006 Слабоспицький О. С.

GitRepos: <https://github.com/MoloZzz/Lab2DataAnalise>