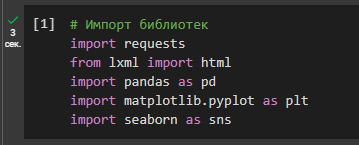
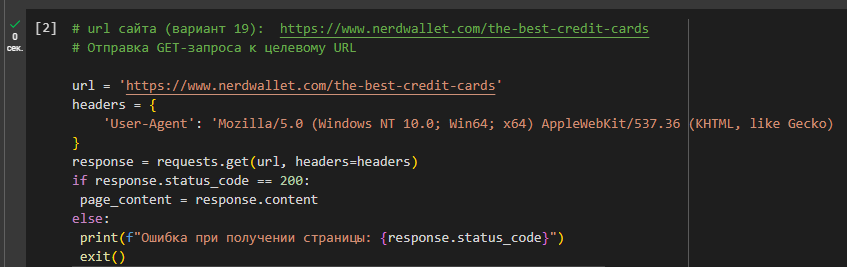
Лабораторная работа №2. Парсинг HTML. XPath+Selenium

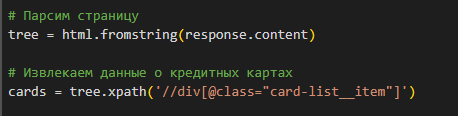
Мареева М.И. Группа БД-231-м

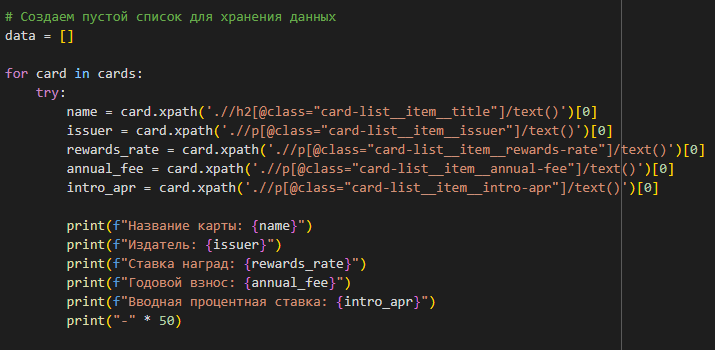
Вариант 19: https://www.nerdwallet.com/best/credit-cards/rewards

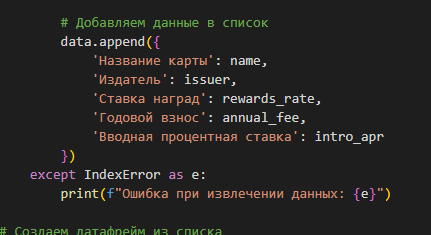
Сначала пробуем использовать XPath в Colab. (полный блокнот расположен на github)

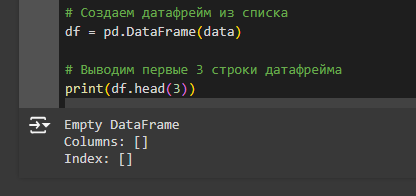












Получаем пустой датафрейм, что могло случиться вследствие того, что таблица на сайте загружается динамически через JavaScript после загрузки основной страницы. При прямом запросе HTML через requests данные могут не быть часть HTML.

Поэтому пробуем получить данные с помощью Selenium, который будет имитировать работу браузера и открытия её человеком.

Установка Selenium:

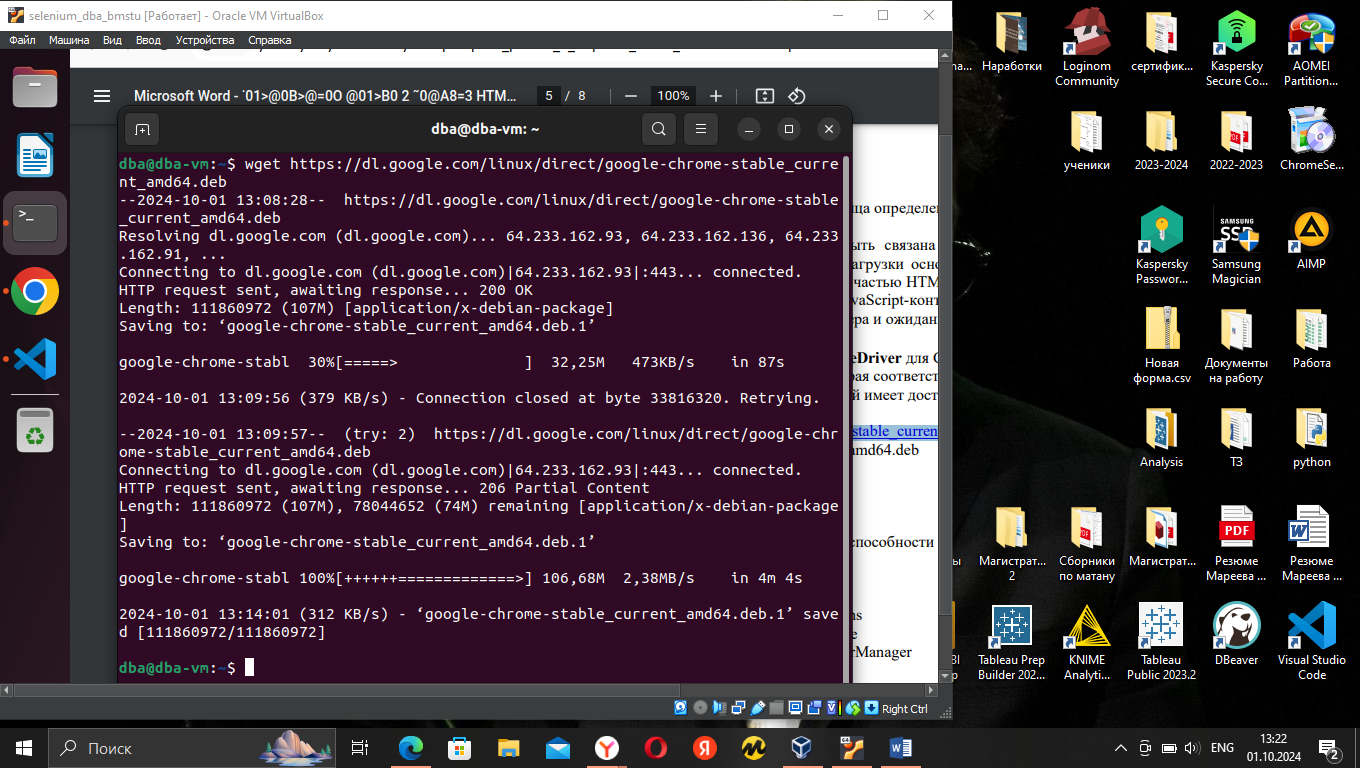
wget https://dl.google.com/linux/direct/google-chrome-stable\_current\_amd64.deb

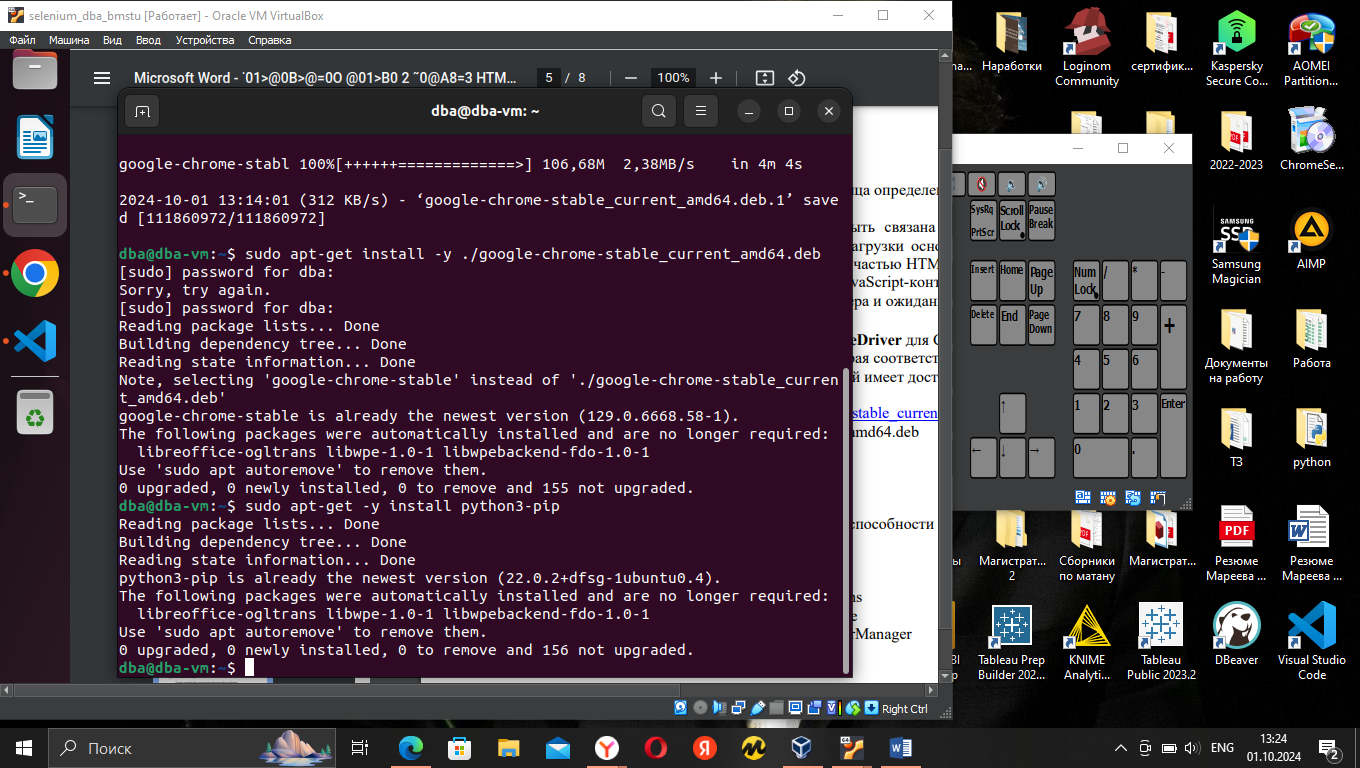
sudo apt-get install -y ./google-chrome-stable\_current\_amd64.deb

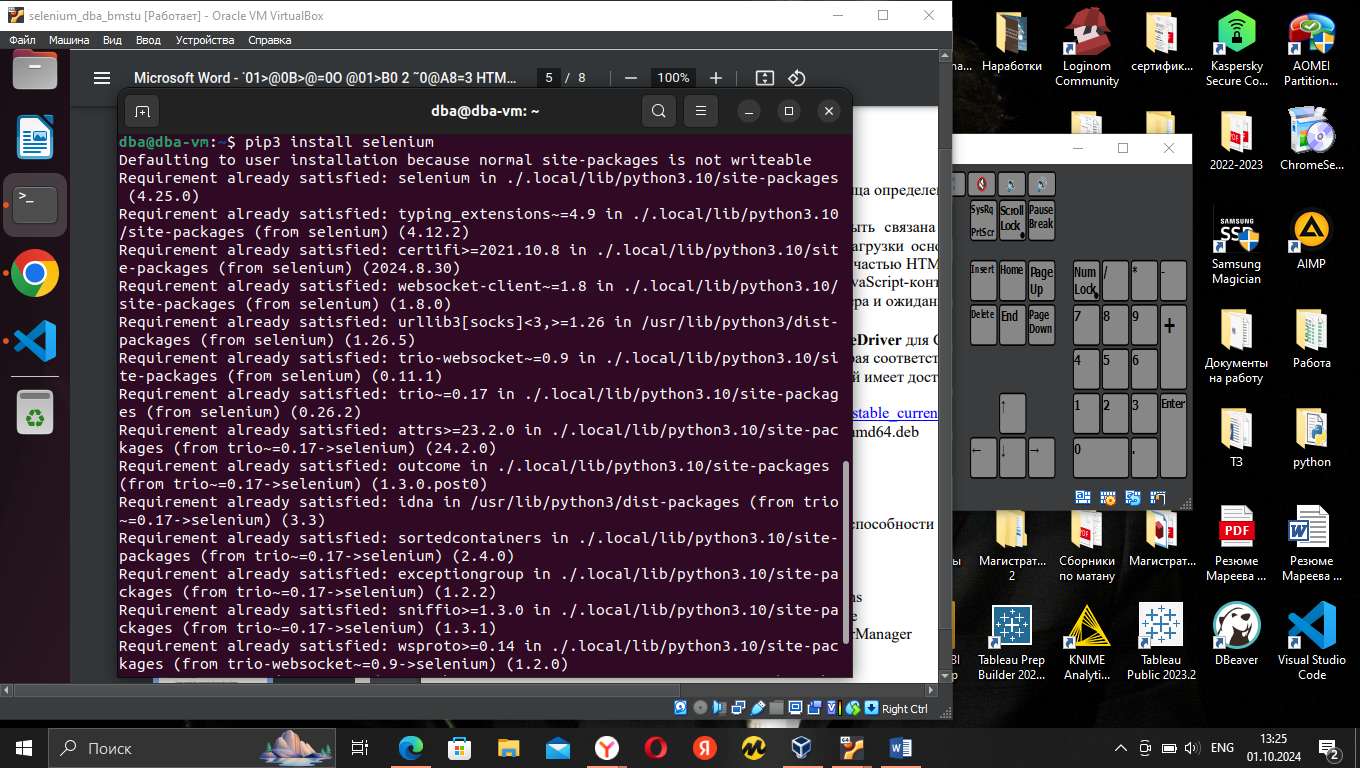
sudo apt-get -y install python3-pip

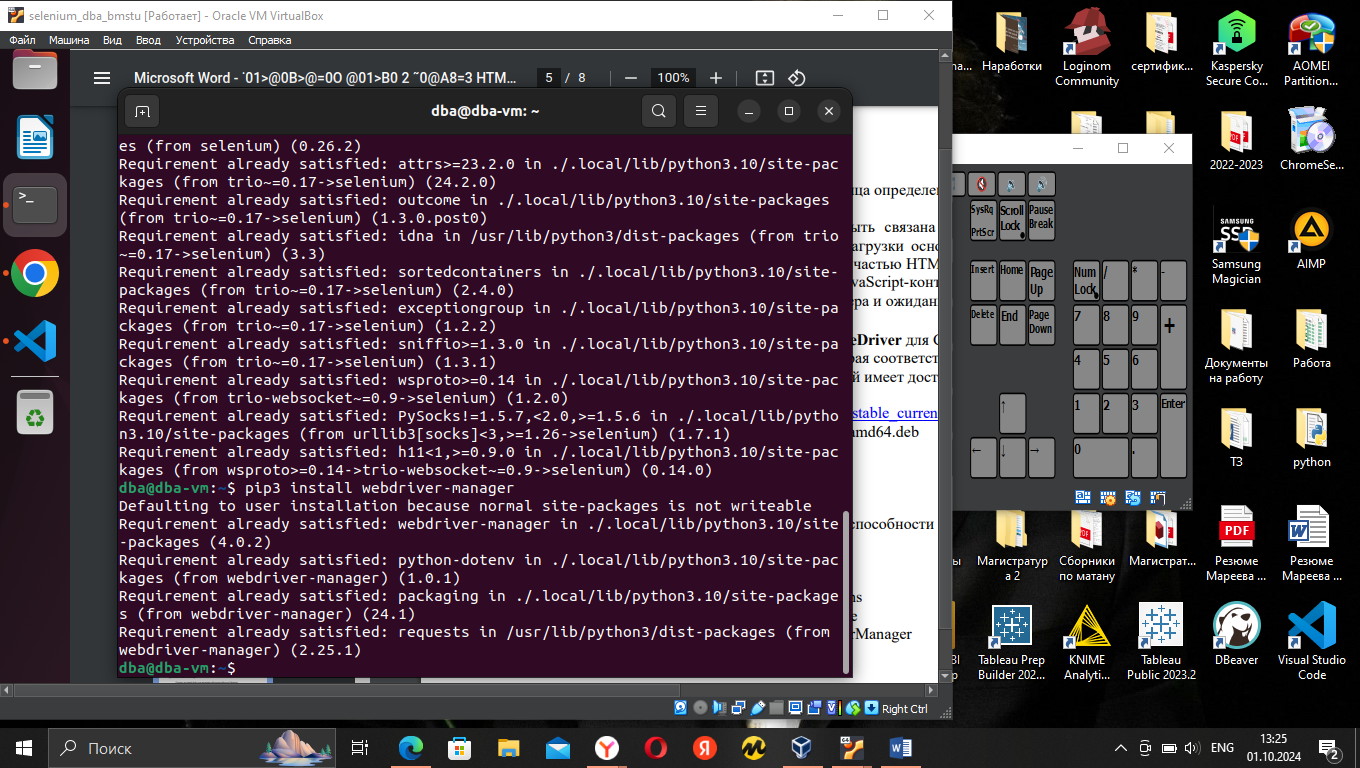
pip3 install selenium

pip3 install webdriver-manager



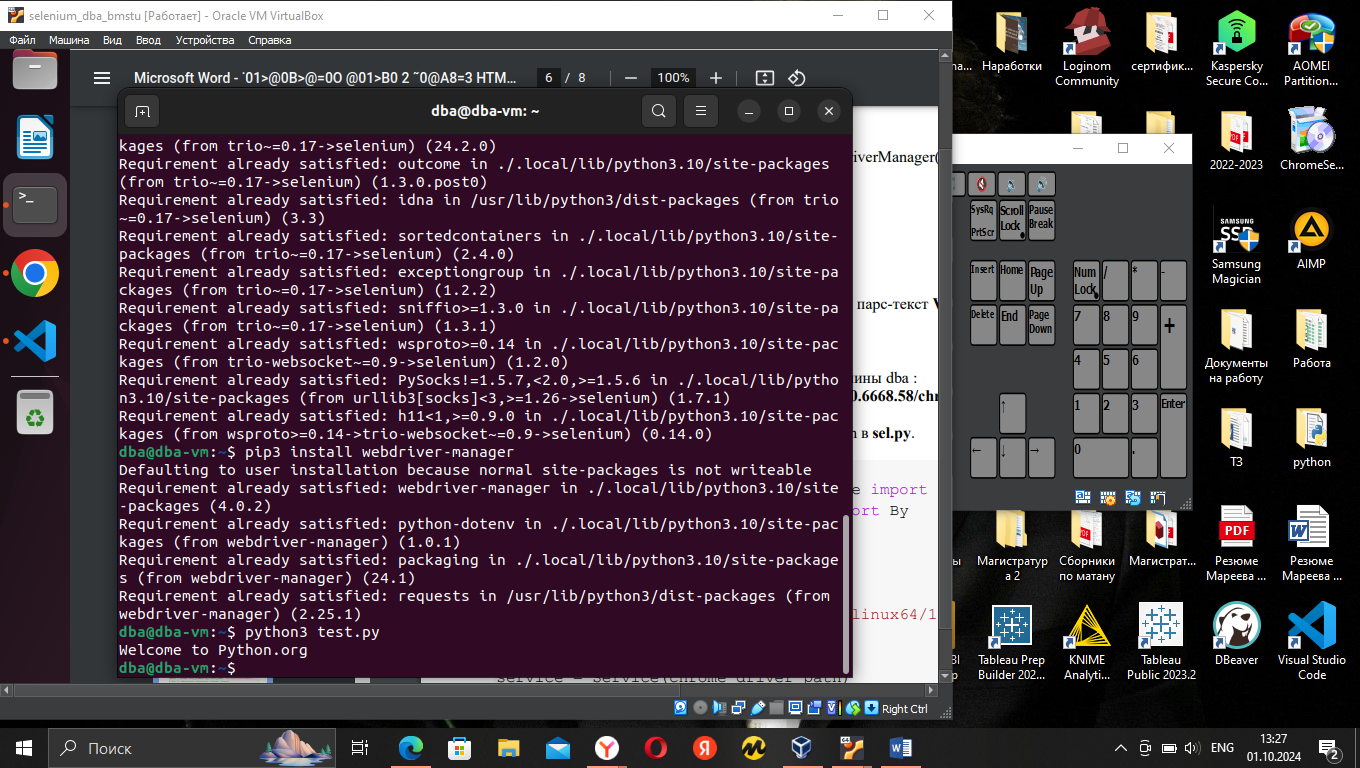






Проверка работы с помощью тестового файла:

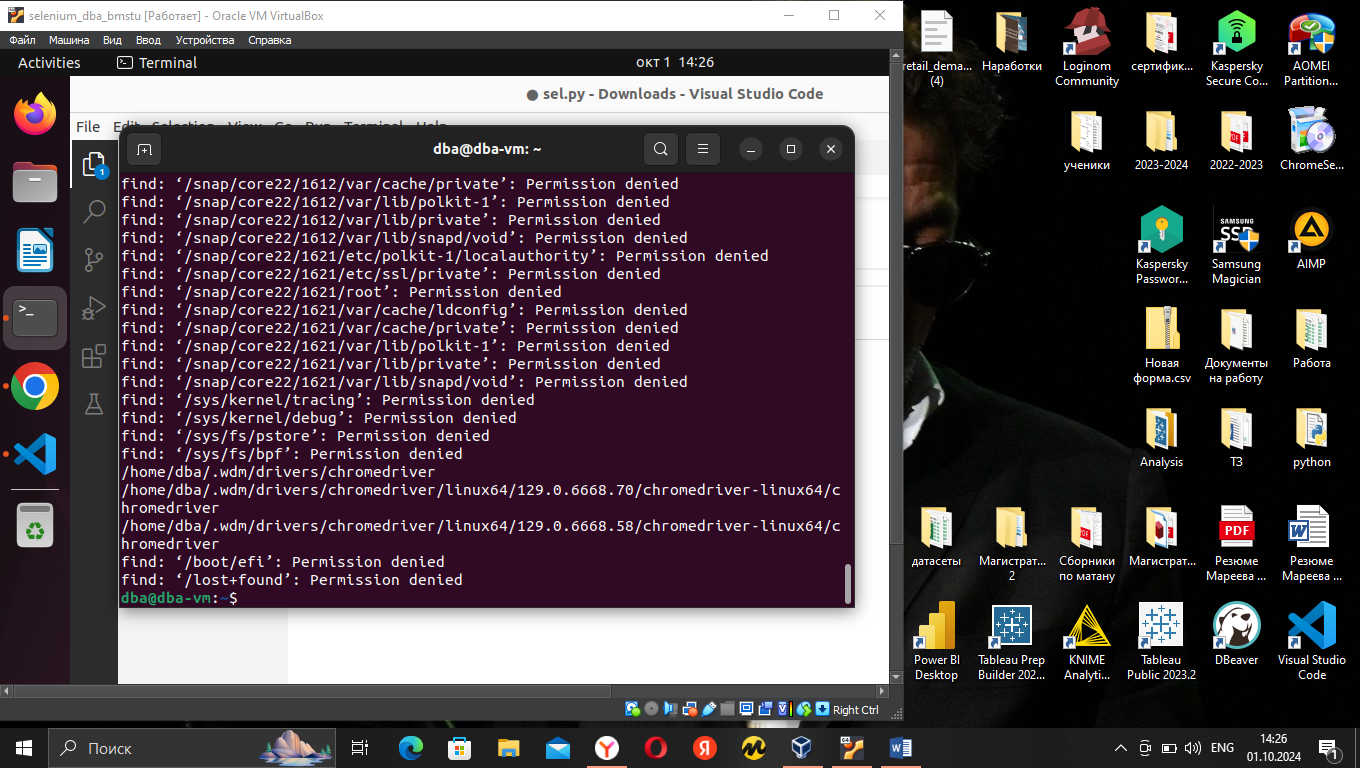
python3 test.py



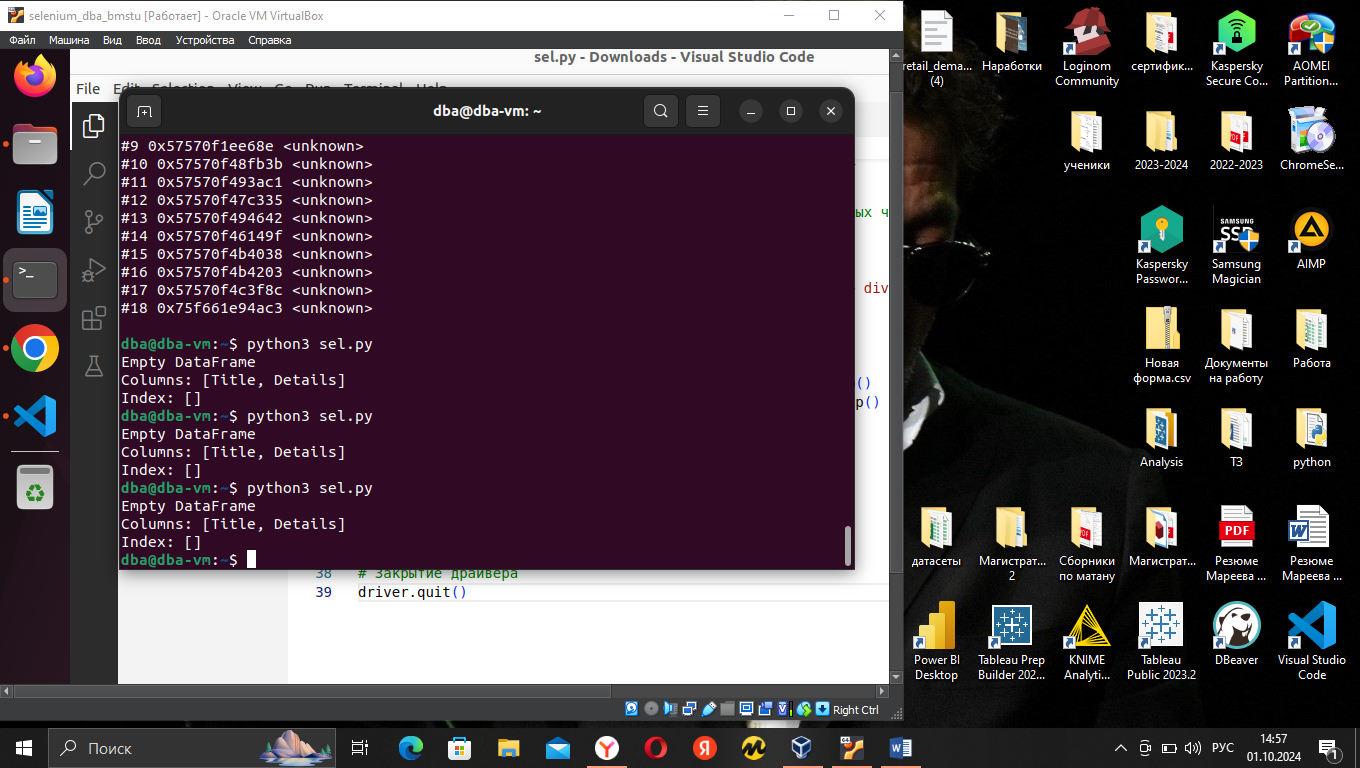
Успешно!

find / -name "chromedriver"

В поиске находим путь для виртуальной машины dba : /home/dba/.wdm/drivers/chromedriver/linux64/129.0.6668.58/chromedriverlinux64/chromedriver



Создаем запрос для получения данных в новом файле sel.py, перемещаем его в ту же папку, что и test.py и запускаем в терминале. (В процессе работы видим, что запрос через селениум самостоятельно открывает и парсит сайт так, как это делает человек)



P.S. я пробовала много раз изменить запрос, но в результате получается пустой датафрейм(

Один из запросов:

from selenium import webdriver

from selenium.webdriver.chrome.service import Service

from [selenium.webdriver.common.by](https://selenium.webdriver.common.by/) import By

import pandas as pd

import time

# Укажите путь к chromedriver

chrome\_driver\_path = "/home/dba/.wdm/drivers/chromedriver/linux64/129.0.6668.58/chromedriver-linux64/chromedriver"

# Шаг 1: Настройка драйвера

Selenium service = Service(chrome\_driver\_path)

driver = [webdriver.Chrome](https://webdriver.chrome/)(service=service)

# Шаг 2: Открытие страницы

url = "<https://www.nerdwallet.com/best/credit-cards/rewards>"

driver.get(url)

# Шаг 3: Ждем несколько секунд для загрузки страницы и данных через JavaScript

time.sleep(5)

# Шаг 4: Поиск таблицы с классом 'ms-overflow-x-auto'

table = driver.find\_element(By.CLASS\_NAME, 'MuiTable-root css-ers4nw')

# Шаг 5: Извлечение заголовков таблицы (thead)

headers = []

thead = table.find\_element(By.TAG\_NAME, 'thead')

for th in thead.find\_elements(By.TAG\_NAME, 'th'):

headers.append(th.text.strip())

# Шаг 6: Извлечение строк данных (tbody)

rows = []

tbody = table.find\_element(By.TAG\_NAME, 'tbody')

for tr in tbody.find\_elements(By.TAG\_NAME, 'tr'):

cells = tr.find\_elements(By.TAG\_NAME, 'td')

row = [cell.text.strip() for cell in cells]

rows.append(row)

# Шаг 7: Создание DataFrame

df = pd.DataFrame(rows, columns=headers)

# Шаг 8: Вывод первых нескольких строк DataFrame

print(df)

# Закрытие драйвера driver.quit()