**به نام خدا**

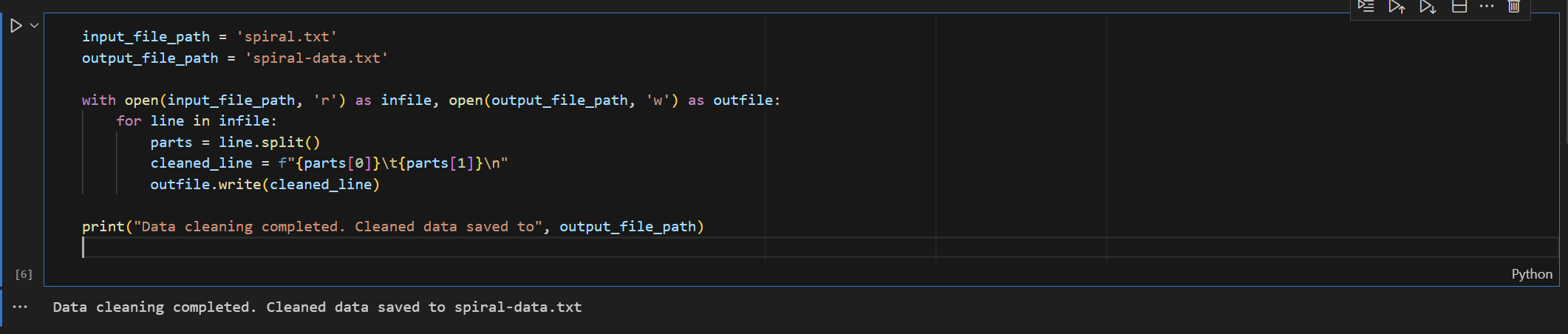
**گزارش پروژه خوشه بندی سلسله مراتبی استاد جوهری**

**دانشجو : محمدرضا شهبازی 401\*\*\*\*\* دانشگاه صنعتی خواجه نصیرالدین طوسی**

در این پروژه طبق اطلاعات درج شده در متن شرح پروژه میخواهیم یک روش تجمعی برای خوشە بندی سلسله مراتبی را به چند روش روی چند مجموعه داده پیادە سازی کنیم.

**بخش اول با استفاده از تابع فاصله ی single link :**

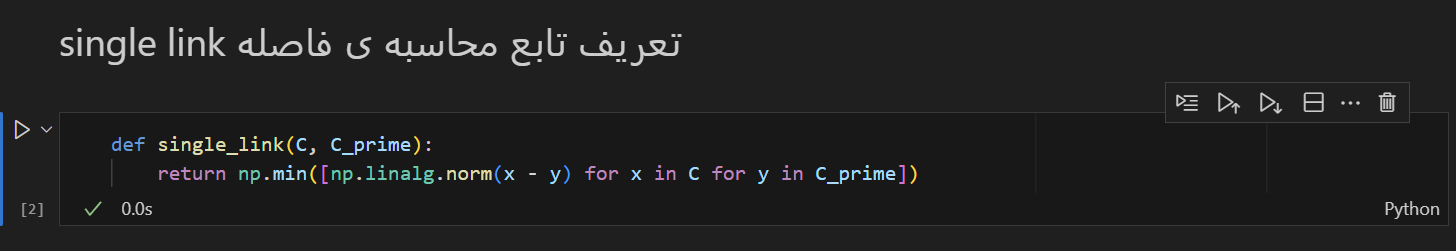
در مرحله ی اول مختصات سوم مجموعه داده هایمان را پاک کردیم چرا که طبق اشاره ی شرح پروژه در این باره به آنها توجهی نداشتیم.  
نمونه فرآیند اجرا شده روی یکی از مجموعه داده ها :



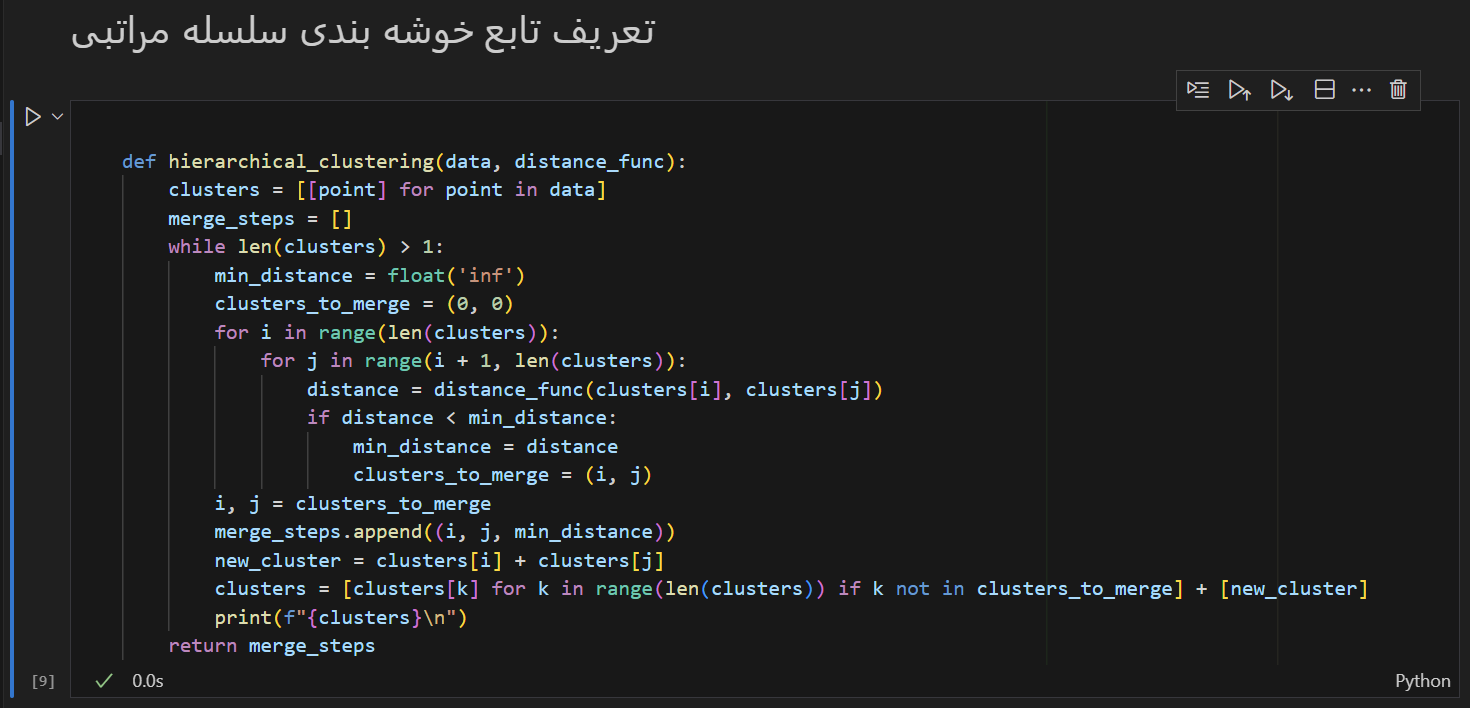
در مرحله ی بعدی مجموعه داده های تصحیح شده را به محیط کد وارد کردیم و از درستی ساختار آن ها مطمعن شدیم :



در این مرحله تابع محاسبه ی سینگل لینک را تعریف کردیم :



سپس در مرحله ی بعد تابع خوشه بندی سلسله مراتبی را تعریف کردیم :



اجرای مراحل خوشه بندی ، نمایش فرم خام دندوگرام خوشه بندی سلسله مراتبی :

در این بخش شما ابتدا خوشه بندی را به صورت خام مرحله به مرحله مشاهده میکنید. همانطور که انتظار میرفت در هر مرحله دو خوشه با یکدیگر ادغام میشوند و در پایان هر مرحله n-1 خوشه باقی میماند.

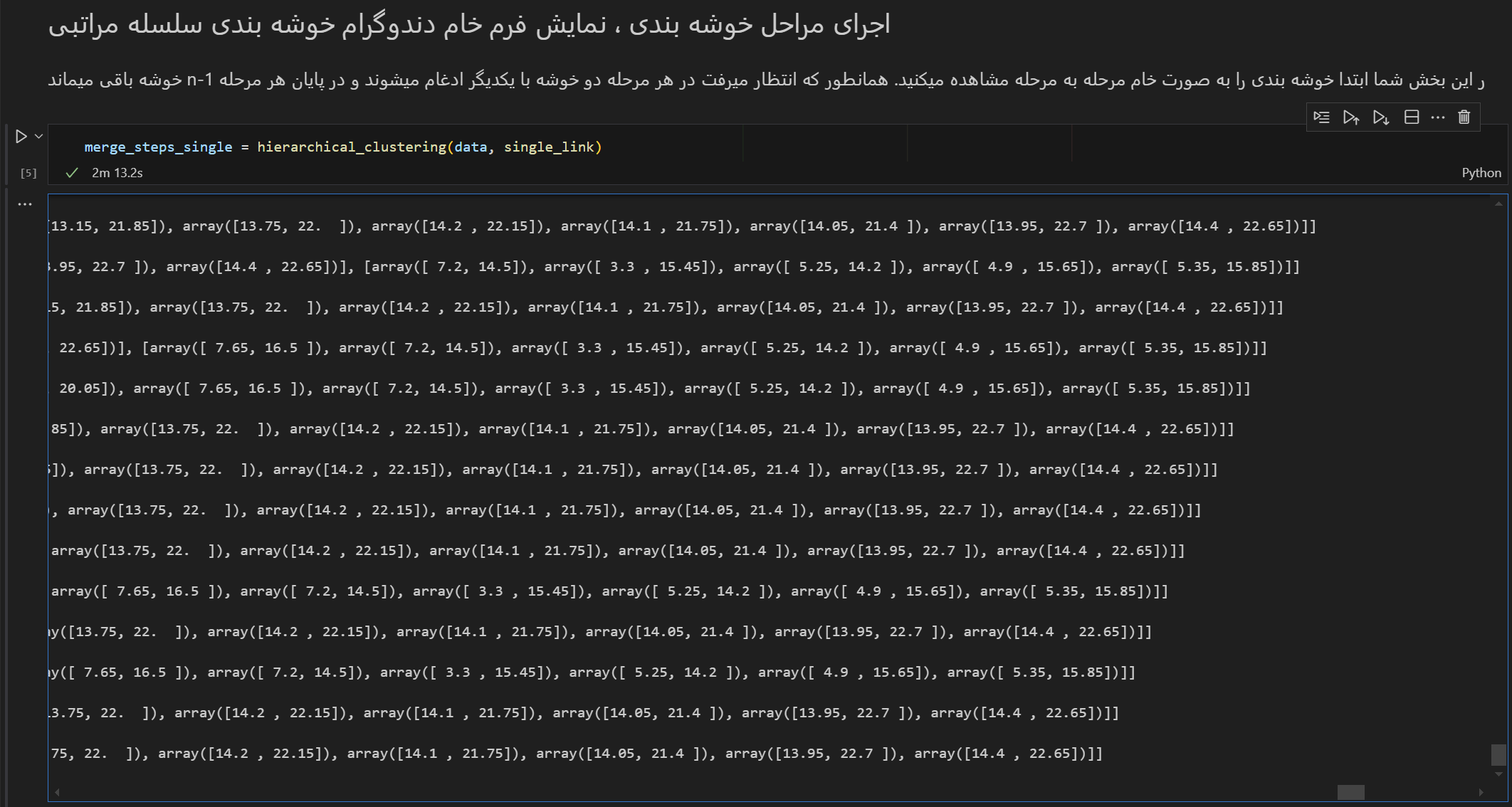
مجموعه داده : Compound.txt در زمان : 3m 32s



مجموعه داده : flame.txt در زمان : 50s



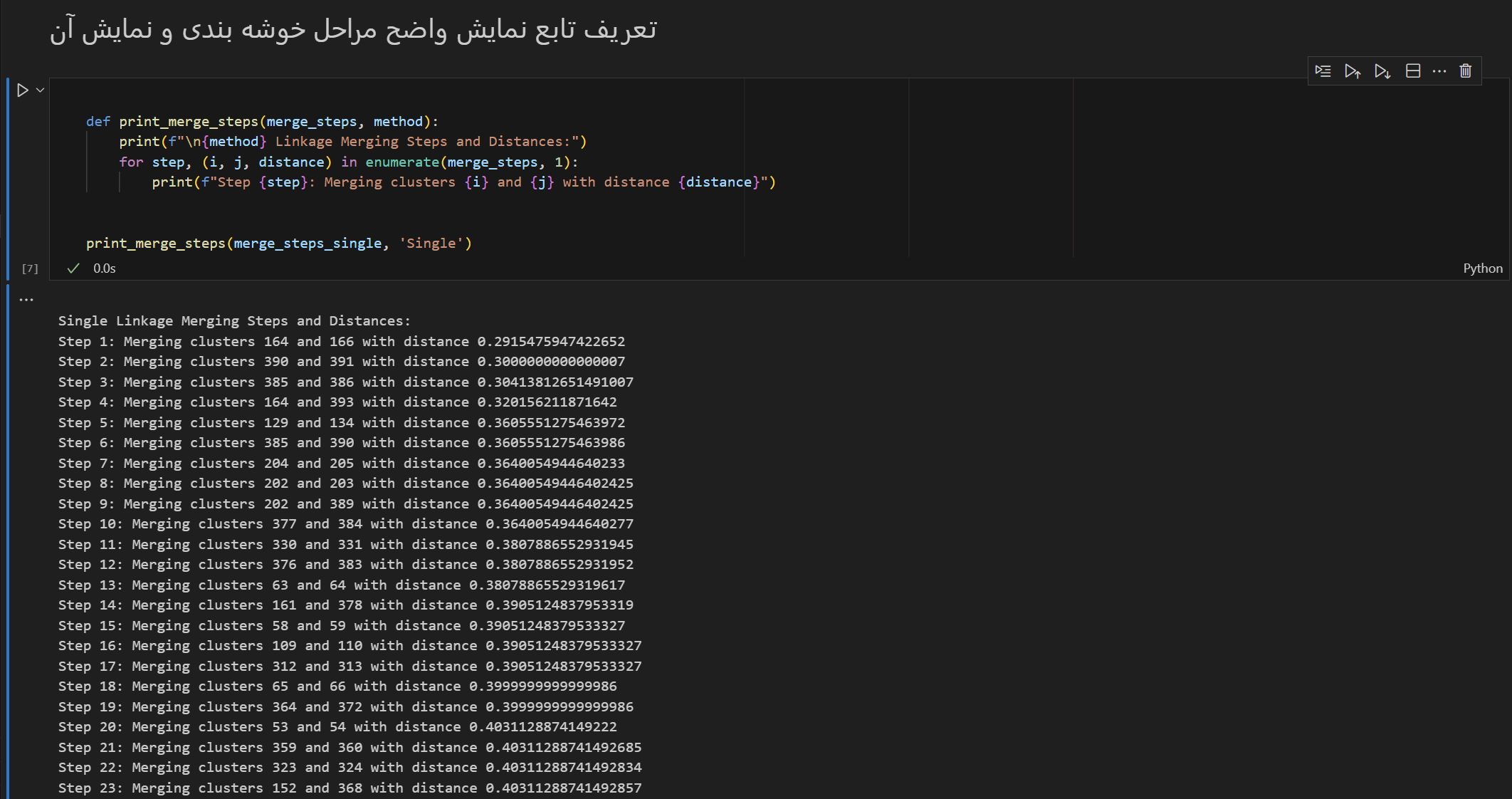
مجموعه داده : jain.txt در زمان : 2m 13s



مجموعه داده : spiral.txt در زمان : 1m 22s



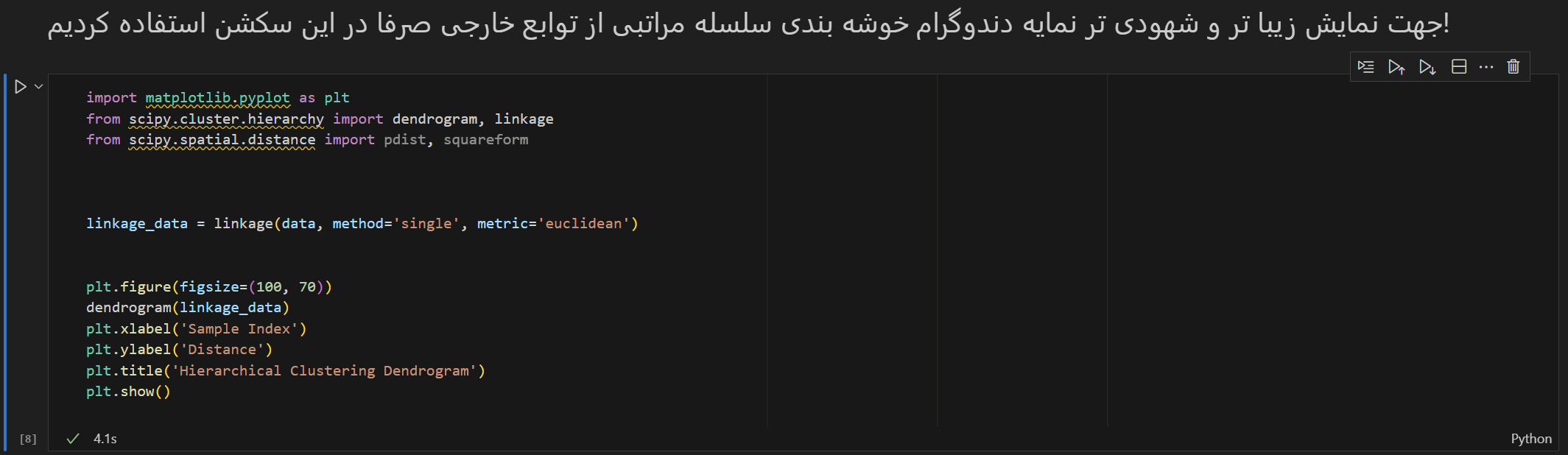
در این مرحله تابع نمایش واضحی تعریف کردیم که بتوانیم بهتر مراحل خوشه بندی را بررسی و نمایش دهیم :



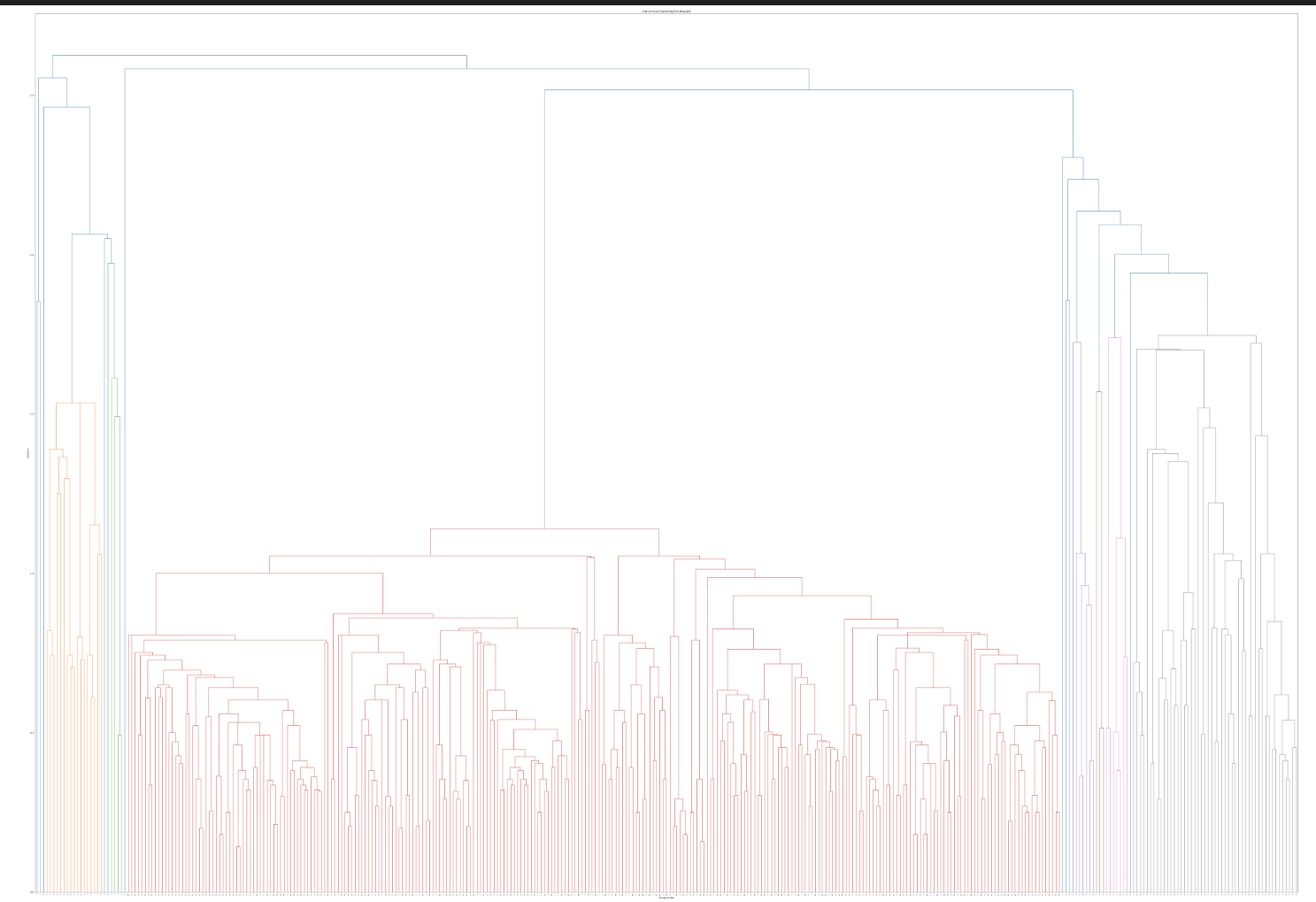
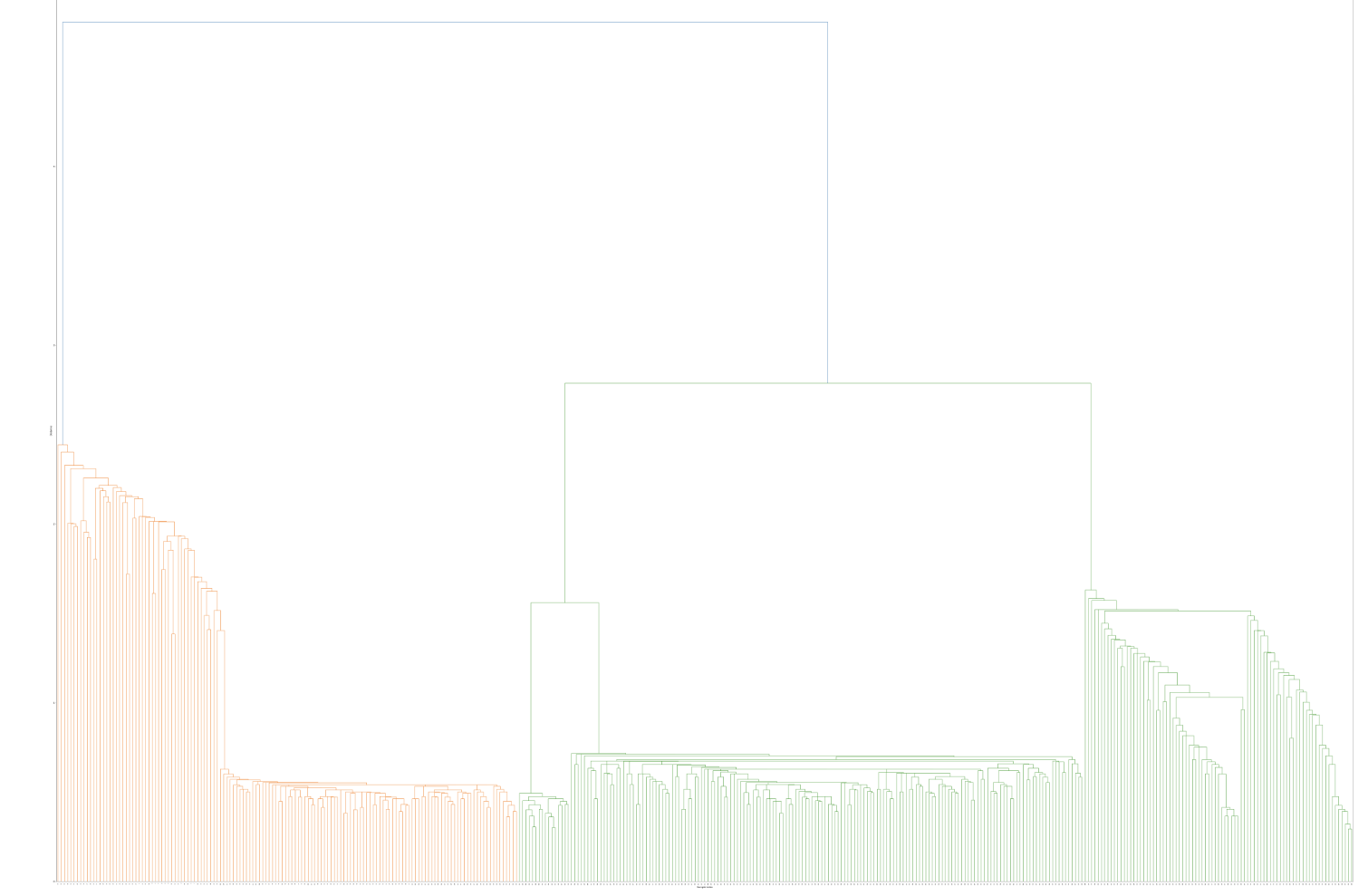
\*توجه : خروجی هر مجموعه داده در فایل نوت بوک مخصوص به خود وجود دارد و به علت طولانی بودن آن ها در اینجا نیاوردیم و به یک نمونه تا مرحله 23 بسنده کردیم.

در مرحله ی بعد صرفا جهت نمایش زیبا تر و شهودی تر نمایه دندوگرام خوشه بندی سلسله مراتبی از توابع خارجی فقط در این سکشن استفاده کردیم!

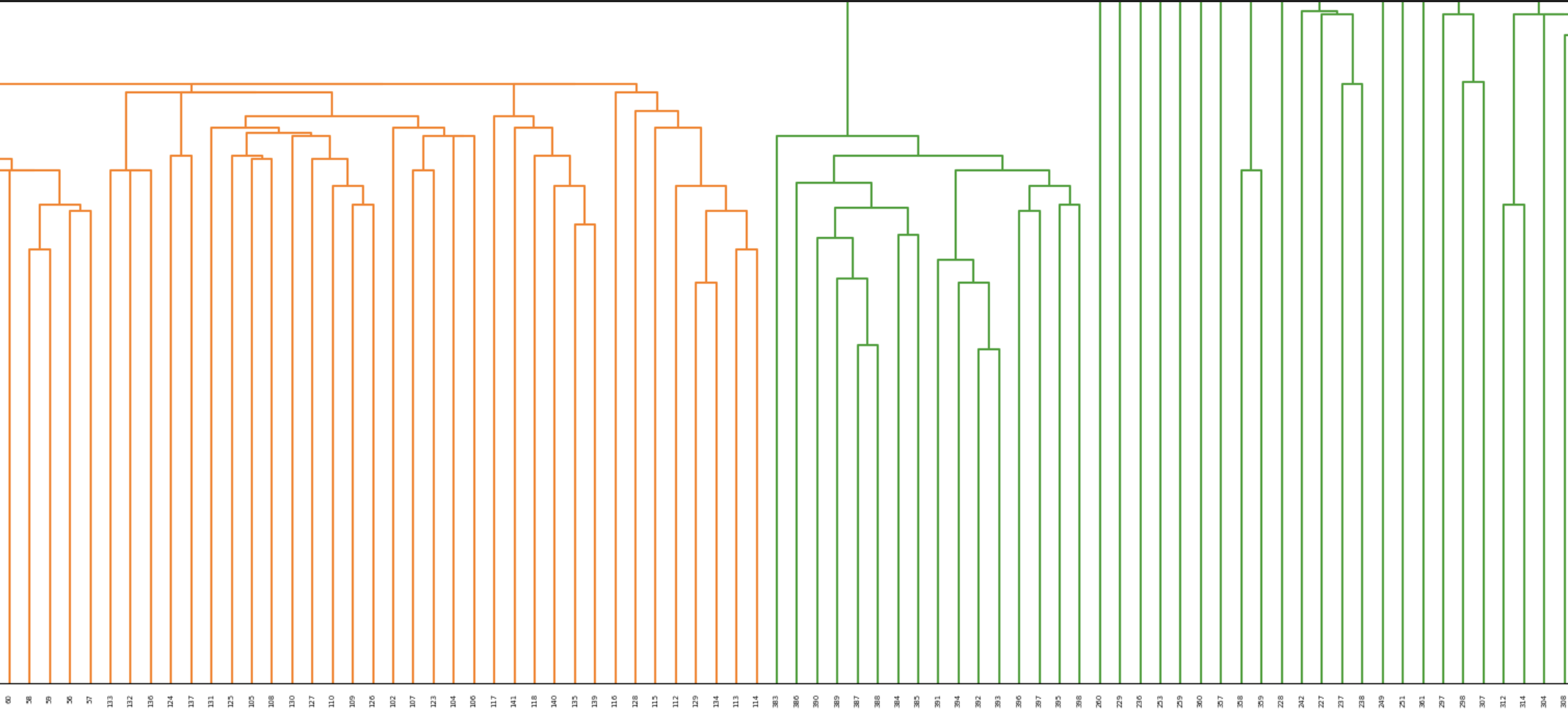
* **تصاویر دندوگرام هر کدام از مجموعه داده ها در فایل نوت بوک قابل زوم کردن و بررسی بهتر میباشد و موجود است. اینجا دو نمونه برای نشان دادن ساختار دقیق دندوگرام آورده ایم.**



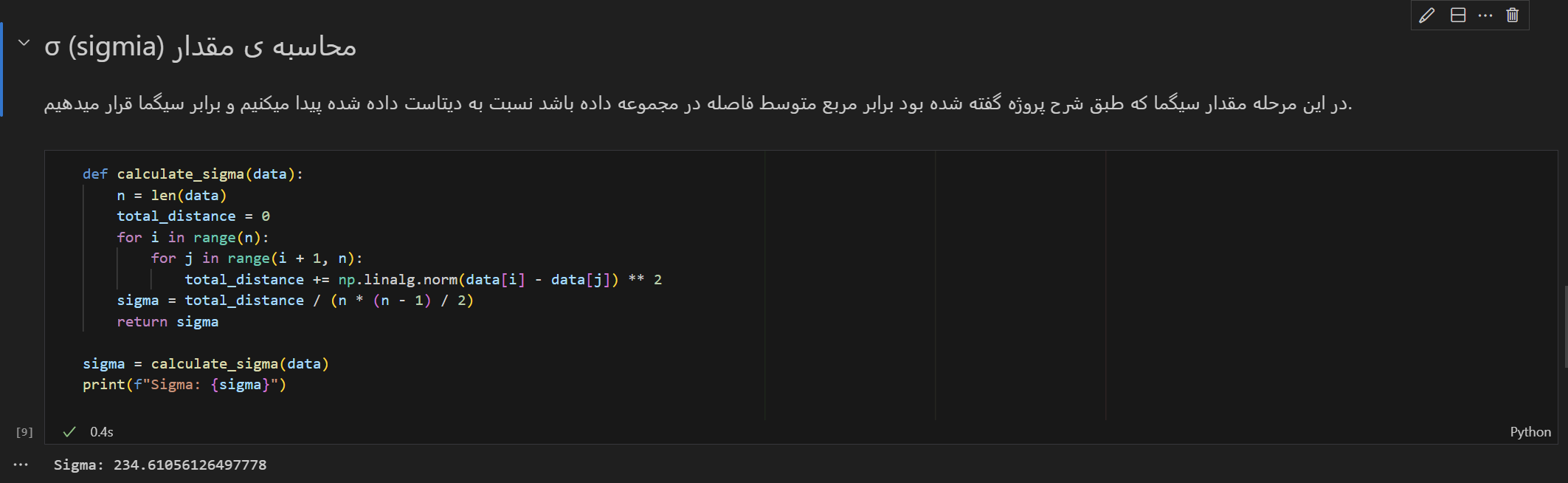
صفحه ی بعد

زوم شده در بخش میانی :

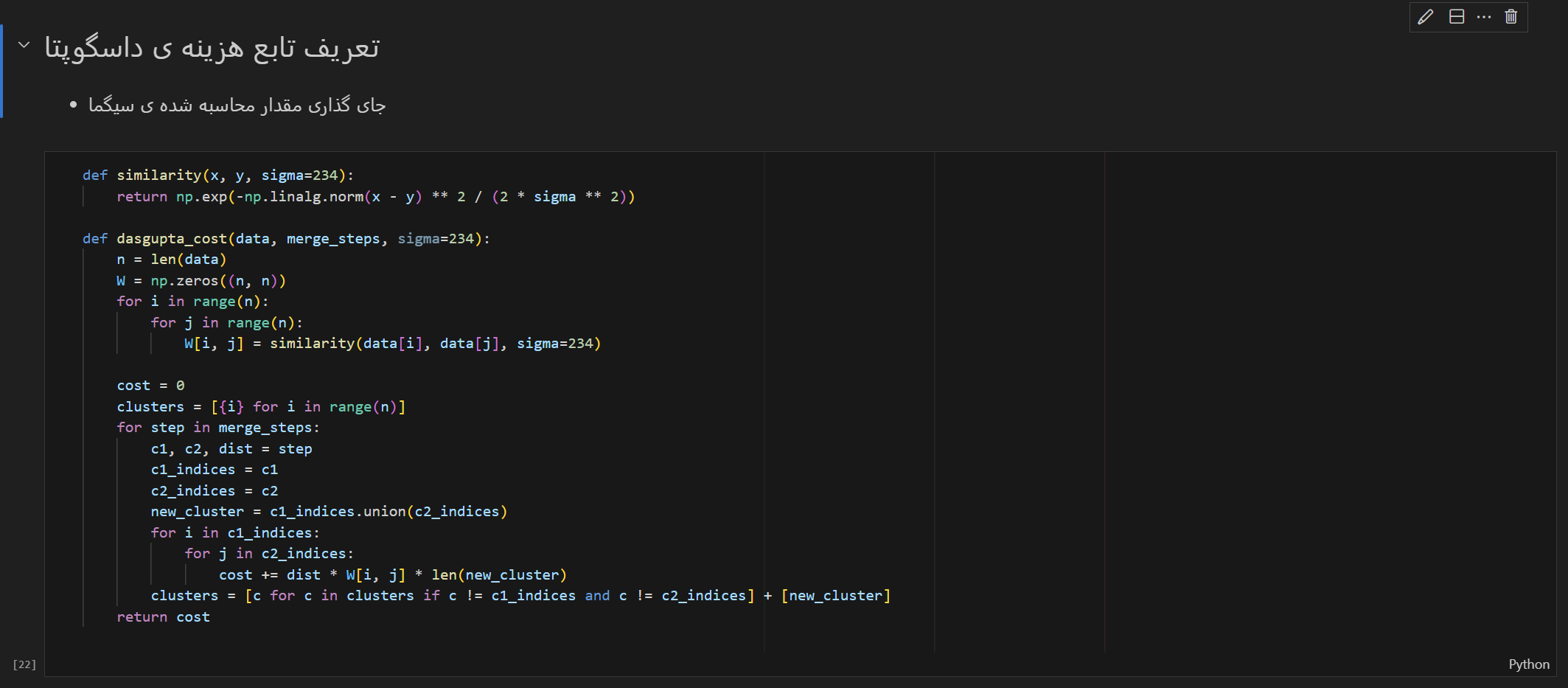


در این مرحله مقدار سیگما که طبق شرح پروژه گفته شده بود برابر مربع متوسط فاصله در مجموعه داده باشد نسبت به دیتاست داده شده پیدا میکنیم و برابر سیگما قرار میدهیم : مورد نمونه : مجموعه ی Compound.txt



**تعریف تابع هزینه ی داسگوپتا**

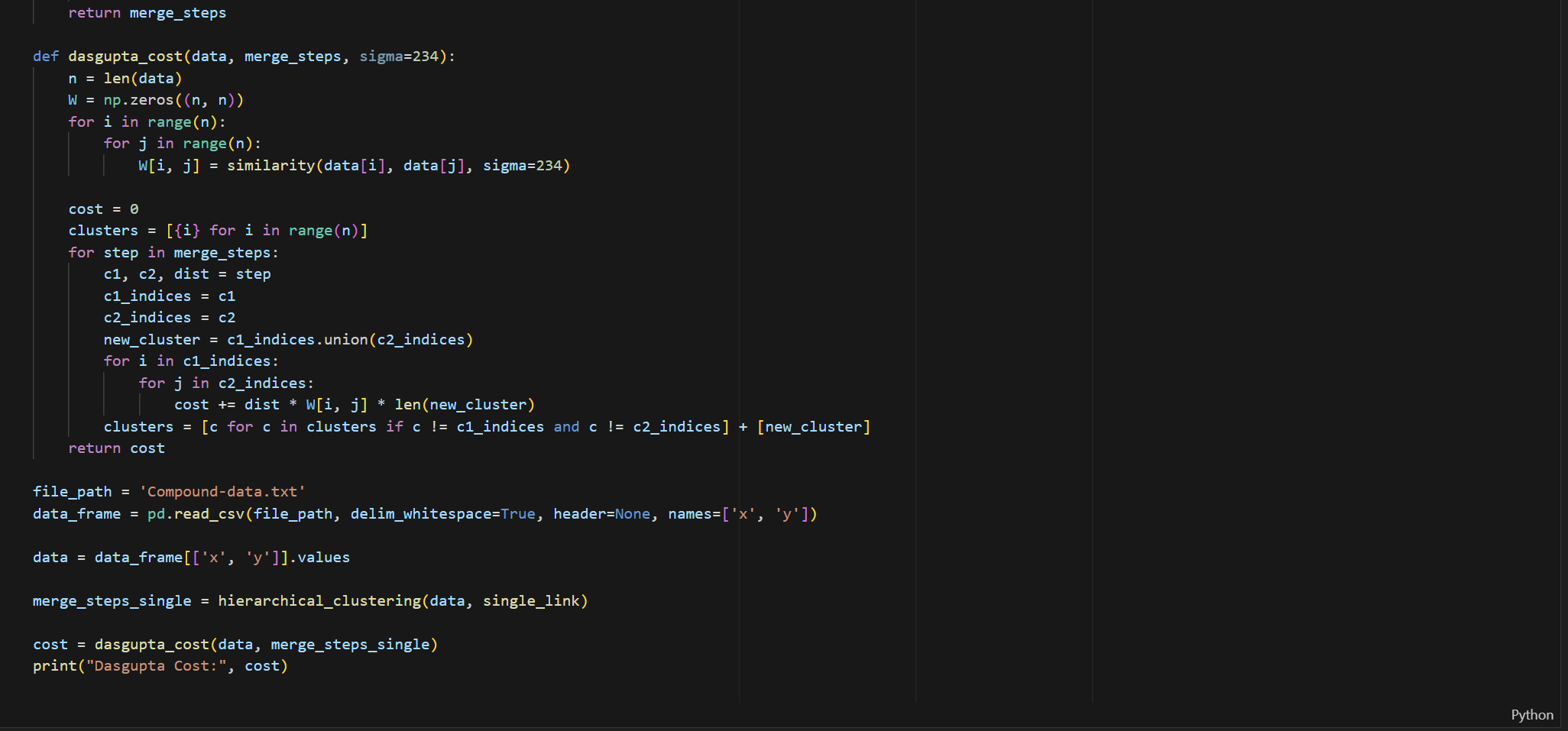
+ جای گذاری مقدار محاسبه شده ی سیگما



**محاسبه ی هزینه ی درخت دو دویی دندوگرام ساخته شده با روش سینگل لینک با تابع داسگوپتایی که تعریف کردیم:**

* در این بخش ما برای رفع ارور های از نوع تایپ ارور که به آن بر میخوردیم مجبور به ایجاد تغییرات جزئی در تابع خوشه بندی سلسله مراتبی مان شدیم تا به جای خروجی هایی از جنس مختصات نقاط ، شماره (عنوان ، اندیس) نقاط را به ما برگرداند لذا در خروجی کد زیر دوباره شاهد مراحل خوشه بندی به فرم خام این بار با اندیس نقاط و در پایان مراحل شاهد مقدار هزینه ی محاسبه شده توسط تابع هزینه ی داسگوپتایمان هستیم.
* صفحه ی بعد

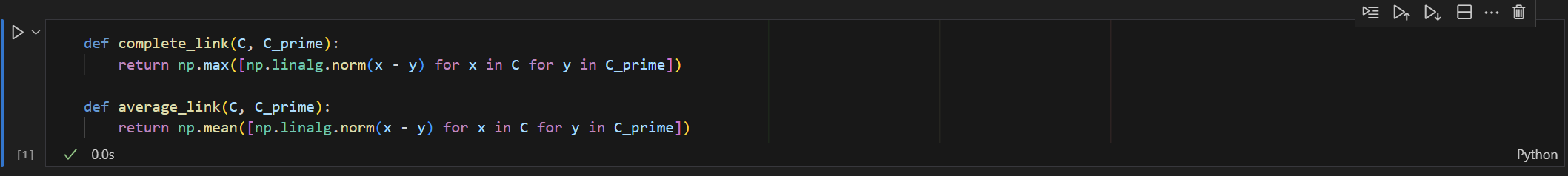






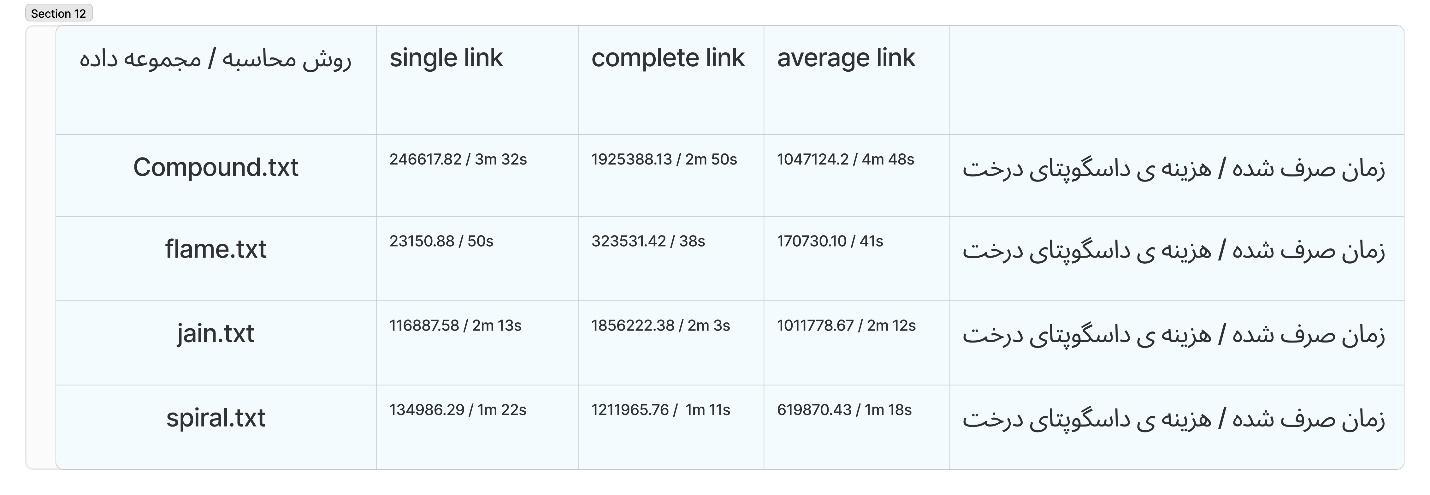
**بخش دوم و سوم با استفاده از تابع فاصله ی complete link و average link :**

در این بخش ها هم تقریبا تمامی مراحل یکسان است به جز بخش تعریف تعوابع محاسبه ی فاصله که در تصویر زیر آورده ایم :



* توجه : تمامی کد ها و خروجی ها به شکل کاملا واضح در 3 بخش single link ، complete link و average link به صورت مرحله مرحله به همراه خروجی های هر بخش در فایل های نوت بوک موجود در فایل پروژه قرار دارد برای بررسی بیشتر قابل مشاهده هستند.

حال به مشاهده ی جدول زمانی و هزینه ی داسگوپتای هر مجموعه داده با 3 روش محاسبه ی فاصله بپردازیم :

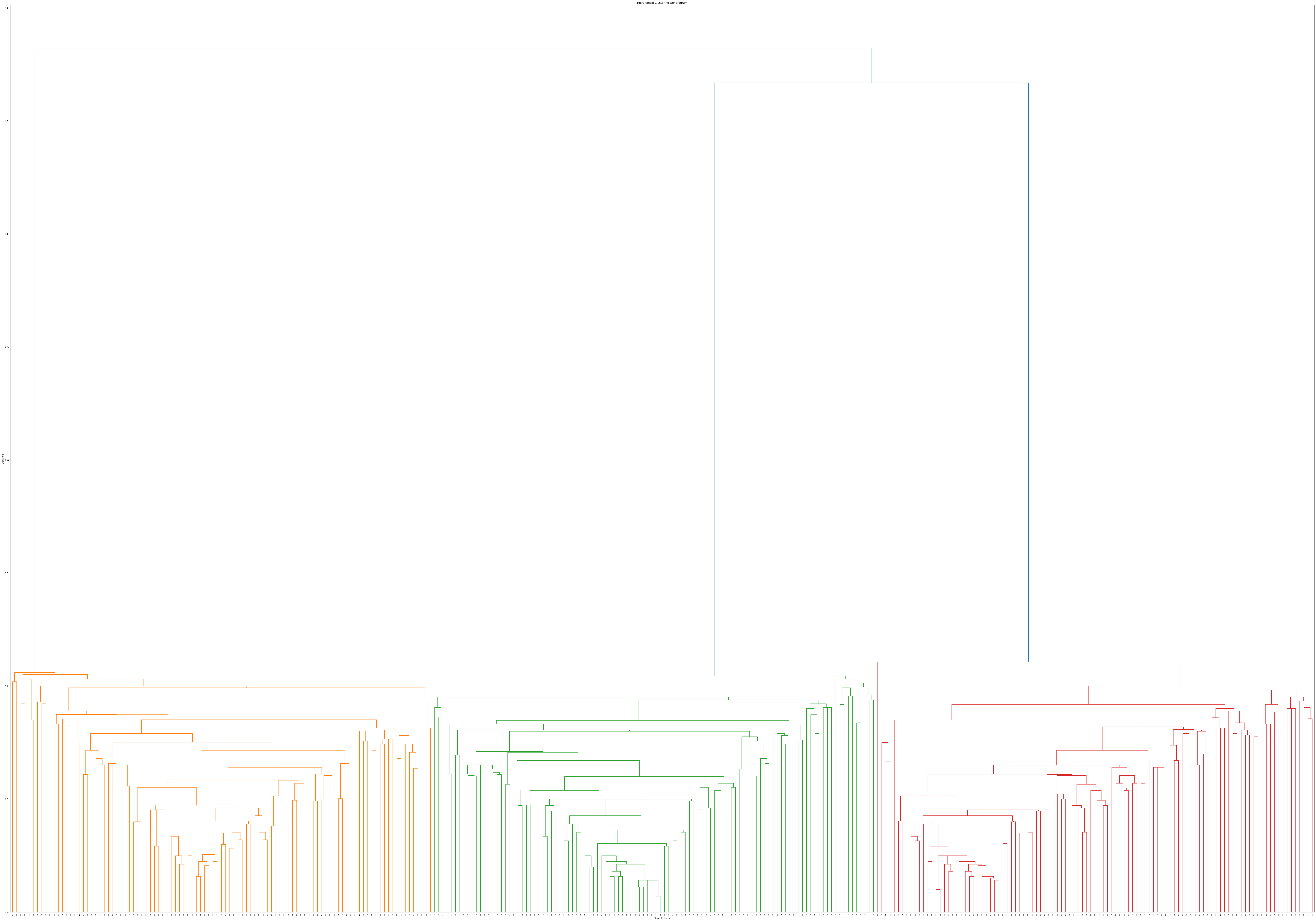


مشاهدات :

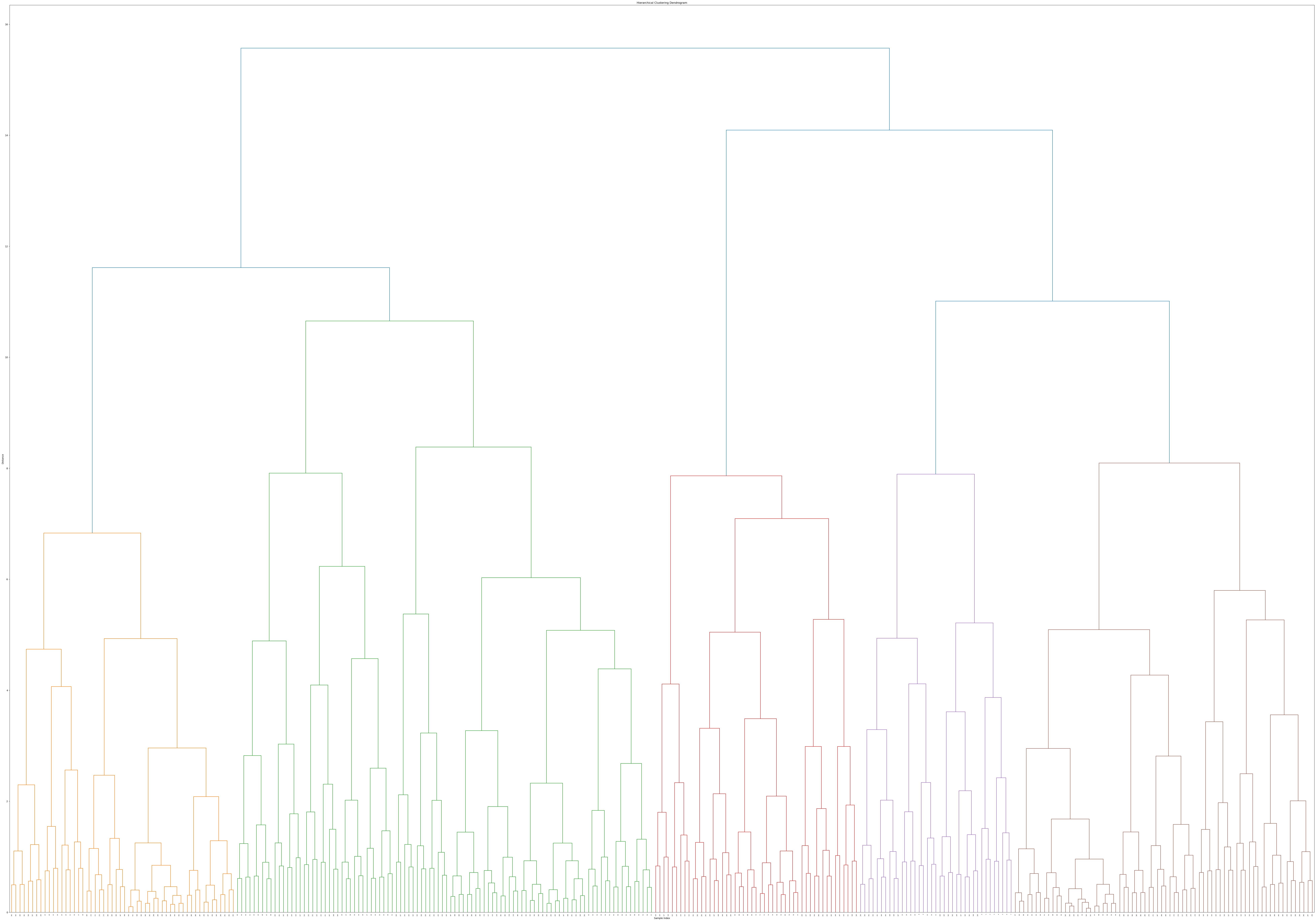
مقدار هزینه داسگوپتا برای درخت دندوگرام تولید شده با روش‌های مختلف خوشه‌بندی سلسله‌مراتبی (single link, complete link, و average link) بسته به ویژگی‌های خاص داده‌ها و ساختار خوشه‌ها می‌تواند متفاوت باشد. با این حال، می‌توان به طور کلی نکاتی را بیان کرد ، روش single link ، طبق مشاهدات از جدول بالا و تصاویر دندوگرام آن تمایل به ایجاد خوشه‌های طولانی و باریک دارد. هزینه داسگوپتا برای این روش معمولاً کمترین مقدار را دارد زیرا تعداد بیشتری از نقاط در برگ‌های مشترک قرار می‌گیرند.   
روش complete link اما تمایل به ایجاد خوشه‌های فشرده‌تر دارد. هزینه داسگوپتا برای این روش معمولاً بیشترین مقدار را دارد زیرا تعداد بیشتری از نقاط در برگ‌های غیرمشترک قرار می‌گیرند. و در آخر روش average link ، روش میانگین‌گیری است و تمایل دارد خوشه‌هایی ایجاد کند که نه به اندازه single link طولانی و باریک هستند و نه به اندازه complete link فشرده. بنابراین هزینه داسگوپتا برای این روش معمولاً بین هزینه‌های دو روش دیگر قرار می‌گیرد.

به عنوان مثال به مشاهده ی دندوگرام مجموعه داده spiral.txt بپردازیم :

Single link



average link



Complete link

