گزارش پروژه سوم امار و احتمال

البرز محمودیان

شماره دانشجویی : ۸۱۰۱۰۱۵۱۴

در این گزارش تنها قسمت های تشریحی را اورده ام که شامل کد نخواهد بود

سوال اول :

۱ـ ا ز کاربرد های اتوانکودرها می توان به مورد زیر اشاره کرد :

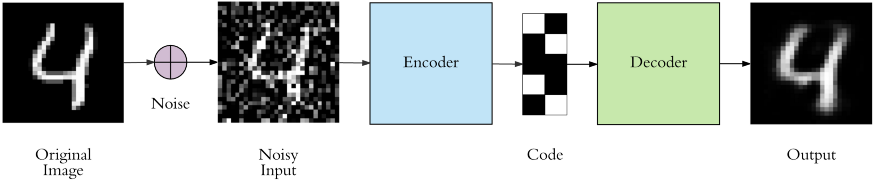
* حذف نویز های تصاویر :

به جای اینکه ورودی و خروجی عینا یکسان باشند، میتوان در هنگام آموزش به وروی نویزی اضافه کرد و در خروجی بدون نویز را از شبکه طلب کرد! در این صورت بعد از آموزش خودرمزگذار را میتوان برای کاربرد حذف نویز استفاده کرد.

می‌توان یک شبکه خودرمزگذار را به منظور یادگیری نحوه حذف نویز از تصاویر، آموزش داد.

* رنگ آمیزی تصاویر
* بازیابی اطلاعات

.



## تشخیص ناهنجاری :

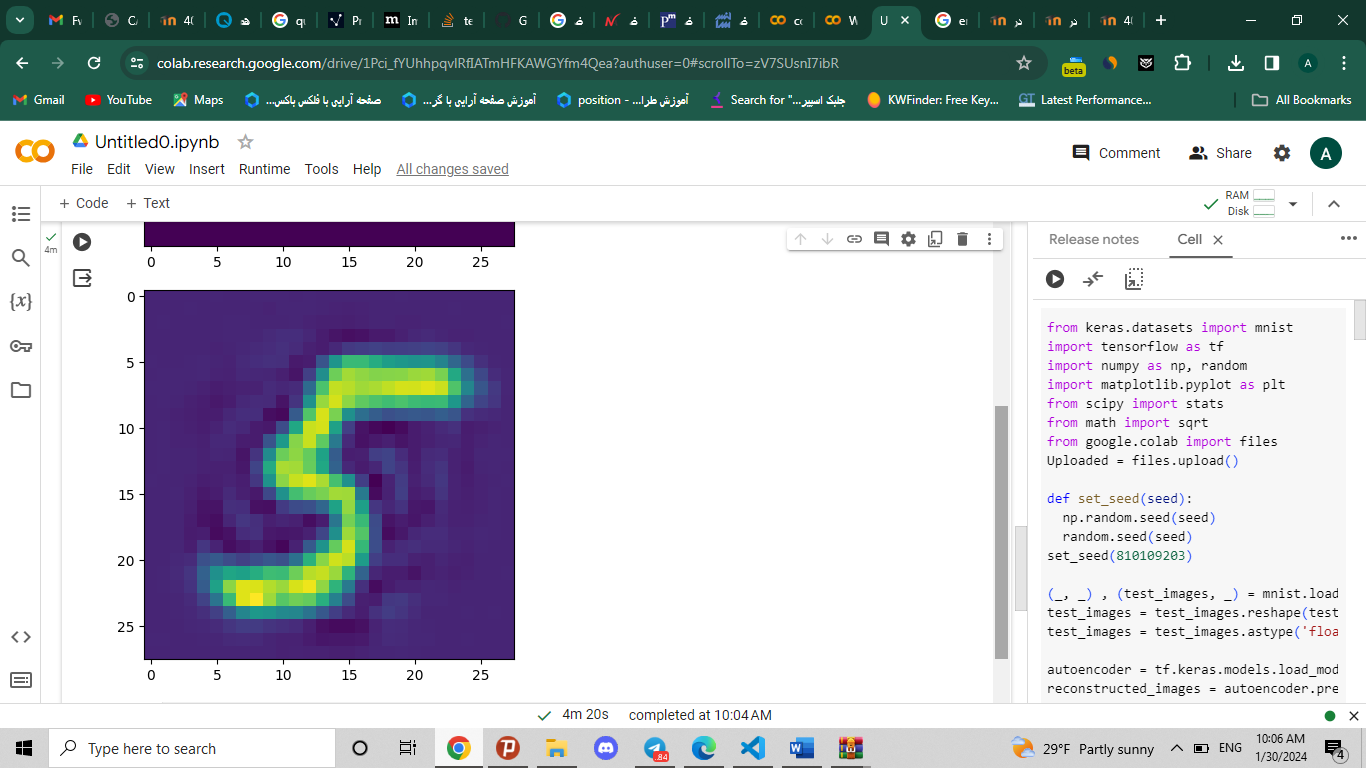
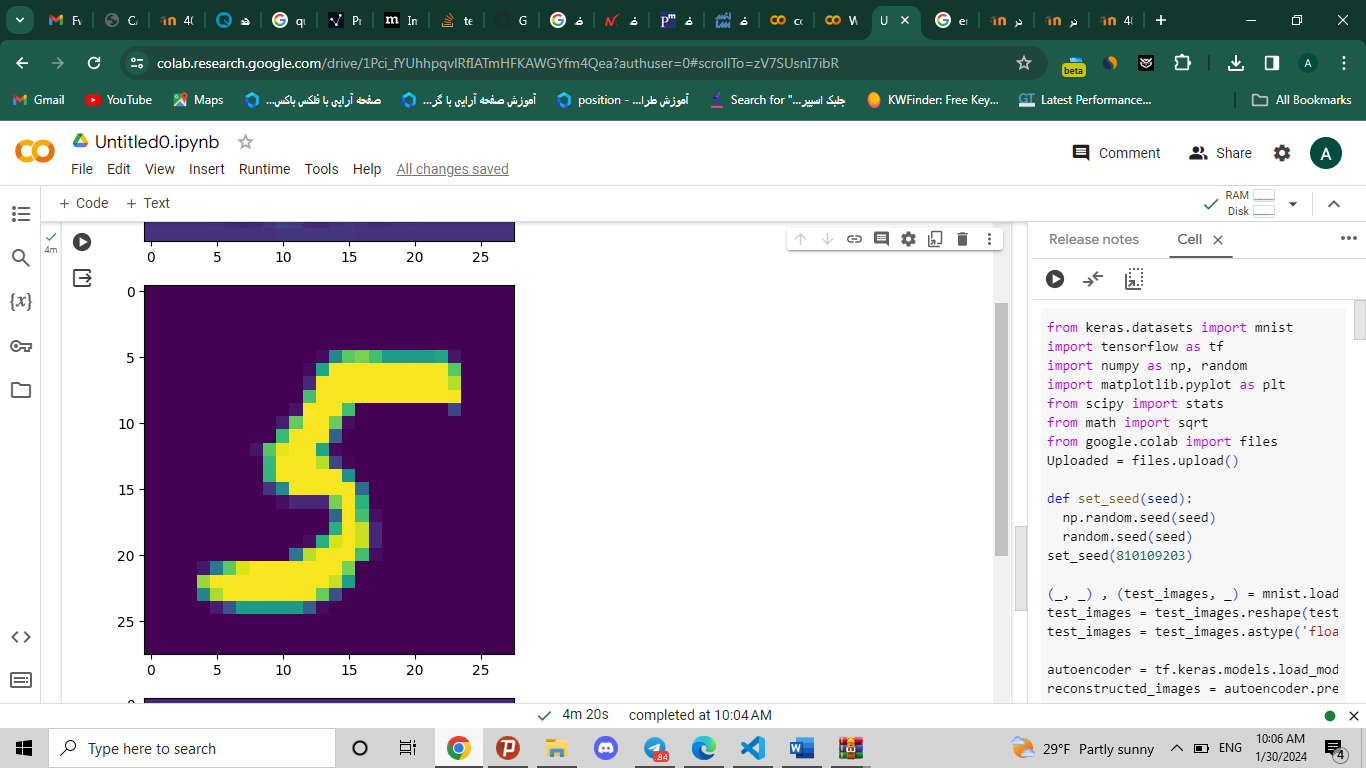
خودرمزگذارها داده هایی که قبلا خود آنها یا مشابه آن ها را در فاز آموزش دیده باشند خیلی بهتر از داده هایی که تا کنون ندیده باشند بازسازی میکنند. با توجه به میزان خطای بازسازی (اختلاف خروجی و ورودی) میتوان فهمید که مشابه آیا این داده قبلا در حین آموزش مشاهده شده یا خیر. در کاربردهای تشخیص ناهنجاری هدف پیدا کردن داده ای است که مشابه داده های قبلی سالم نباشد و در نتیجه میتوان این کابرد را از این شبکه ها انتظار داشت اما وقتی مشاهدات فقط یک دسته اعداد و یک بعدی هستند، تشخیص آن‌ها آسان است، اما وقتی هزاران مشاهده چند بعدی داشته باشید، برای تشخیص این مقادیر به روش‌های هوشمندانه­‌تری نیاز خواهید داشت.

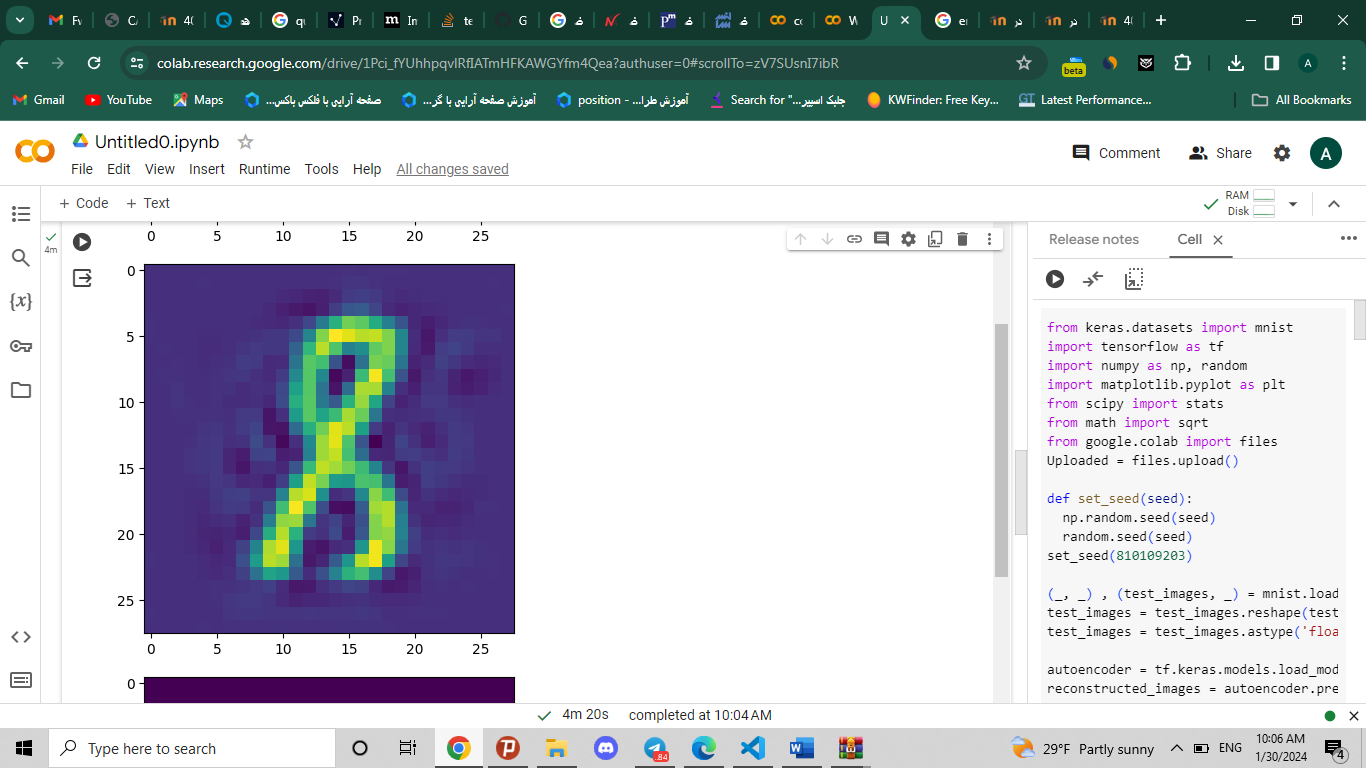
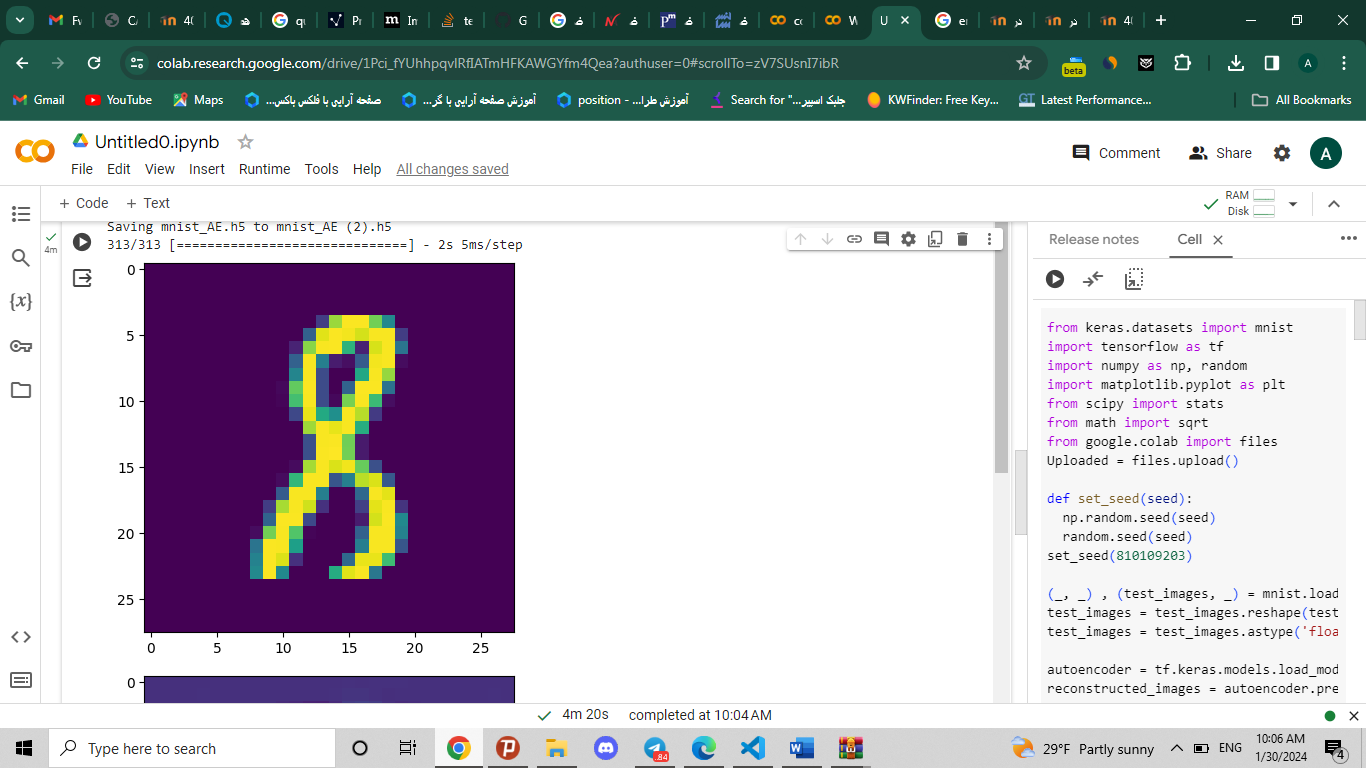
۲ـ

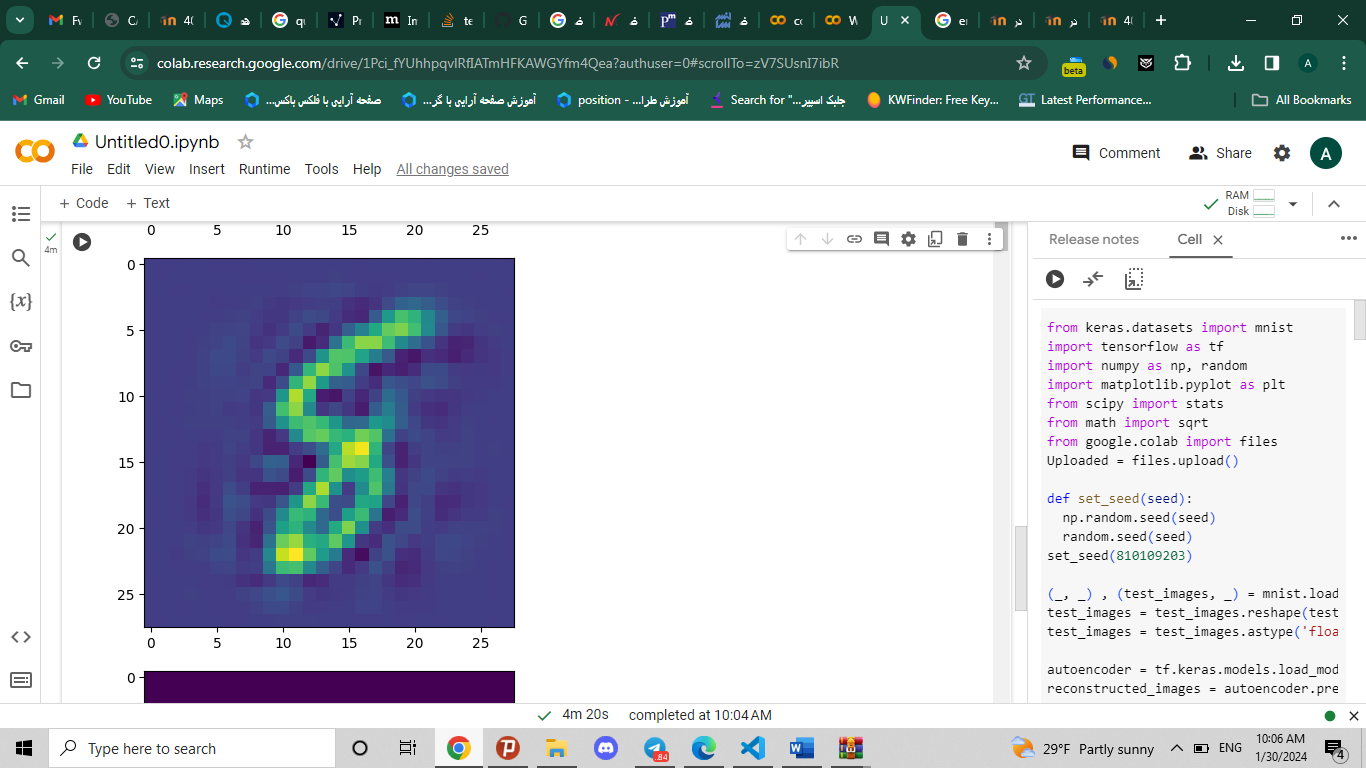
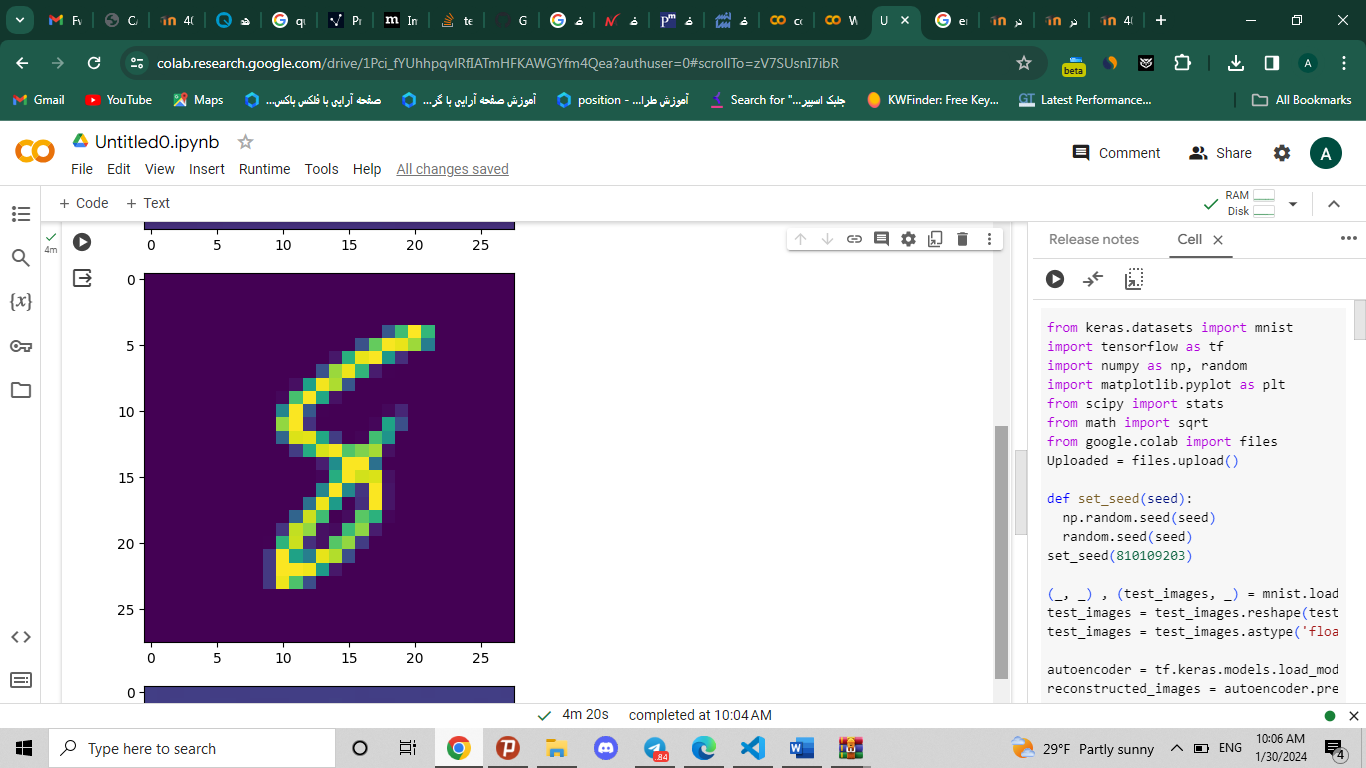
این خطا بخاطر فضای نهان ایجاد می شود. این فضا بسیار تاثیر مهمی در فرایند encode کردن دارد. درواقع اگر ابعاد فضای نهان نسبت به داده ورودی کوچکتر باشد اتوانکودر نمی تواند تمام ویژگی های ورودی را داشته باشد و برعکس اگر این ابعاد بسیار بزرگتر از ورودی باشد نیز امکان خطا وجود دارد چرا که تبدیل ورودی به ابعادی بزرگتر از ابعاد خودش می تواند باعث بروز مشکل شودو در نتیجه میزان خطای اتوانکودر می تواند به ما درباره درستی یا نادرستی ابعاد فضای نهان خبر دهد و باید تعادلی میان ابعاد ورودی و ابعاد این فضای نهان حفظ شود

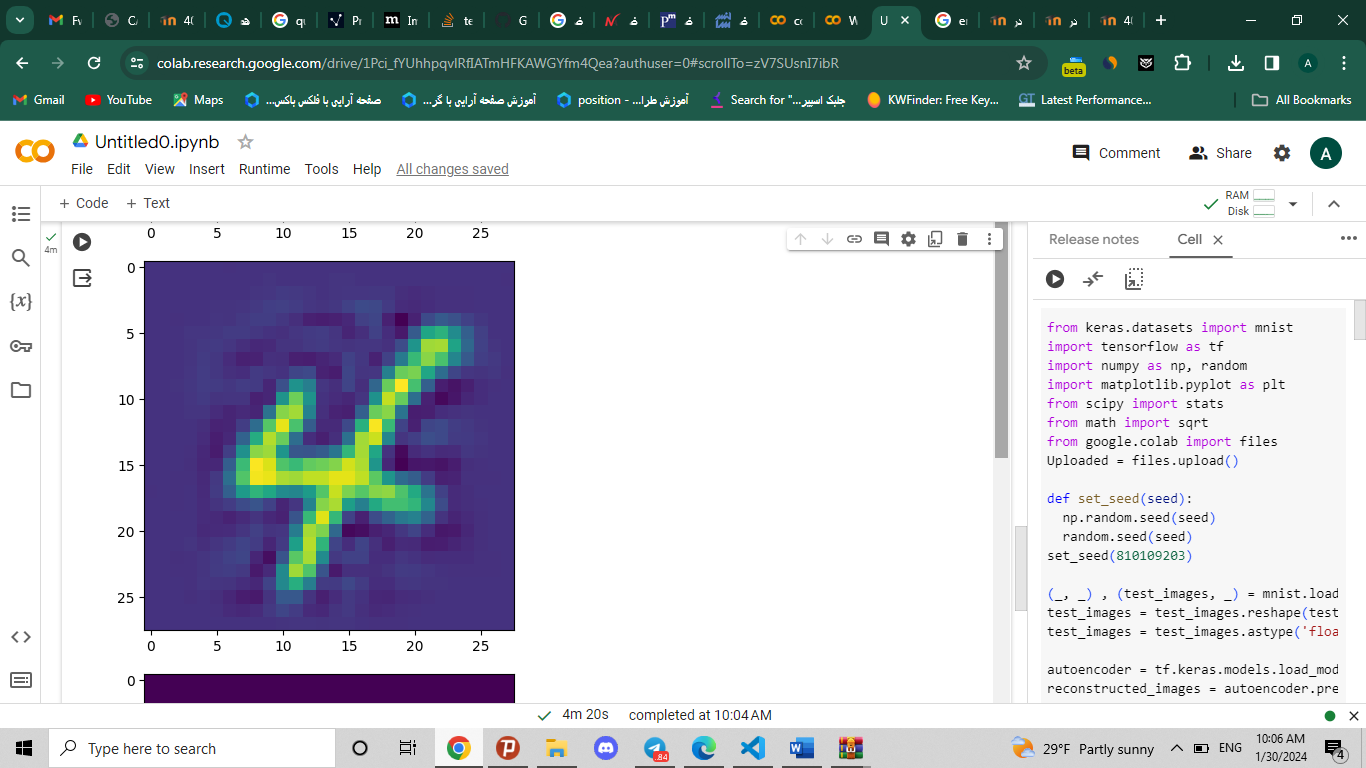
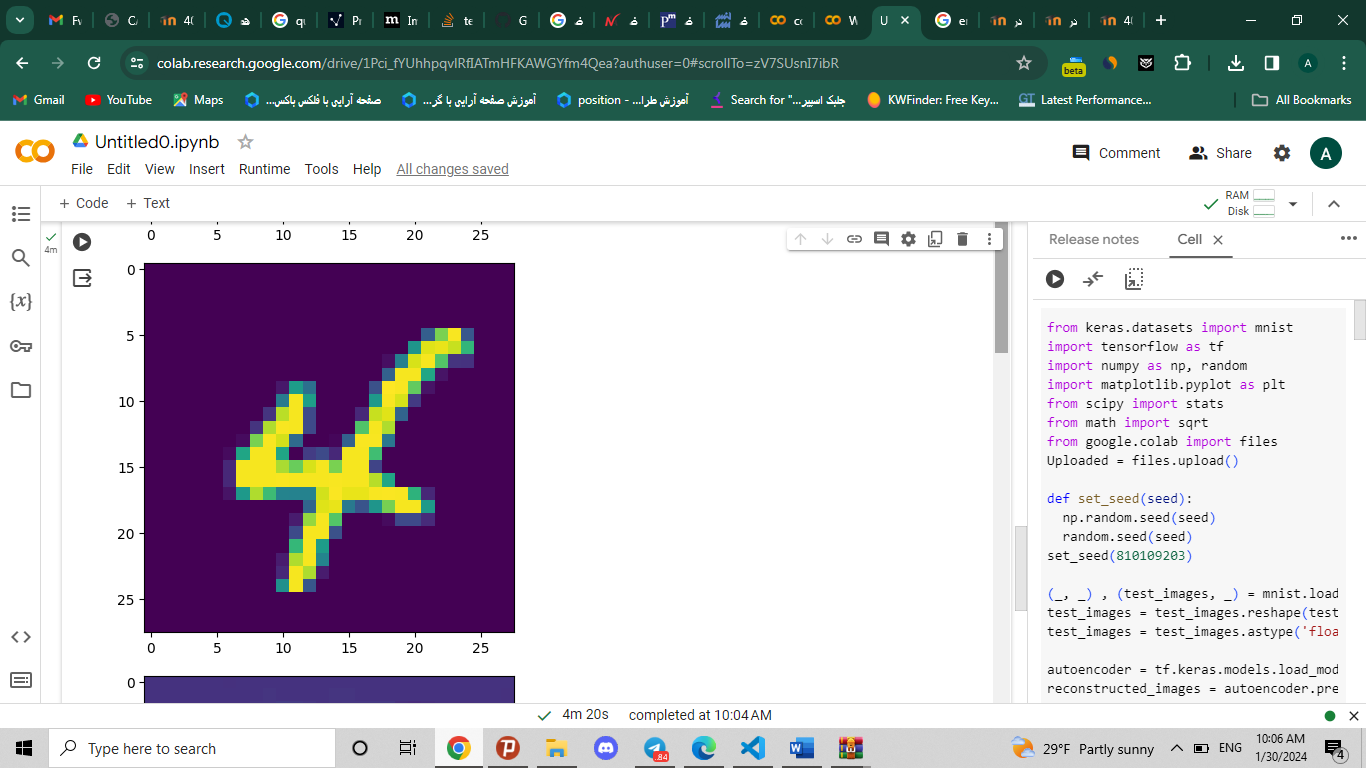
۳ـ ج:

خروجی مانند عکس های زیر خواهد بود و از انجا که درهر بار اجرا عدد ها به صورت رندوم انتخاب می شوند ترتیب اعداد زیر کاملا تصادفی می باشد. که عکس های راست test و عکس های چپ reconstructed هستند.



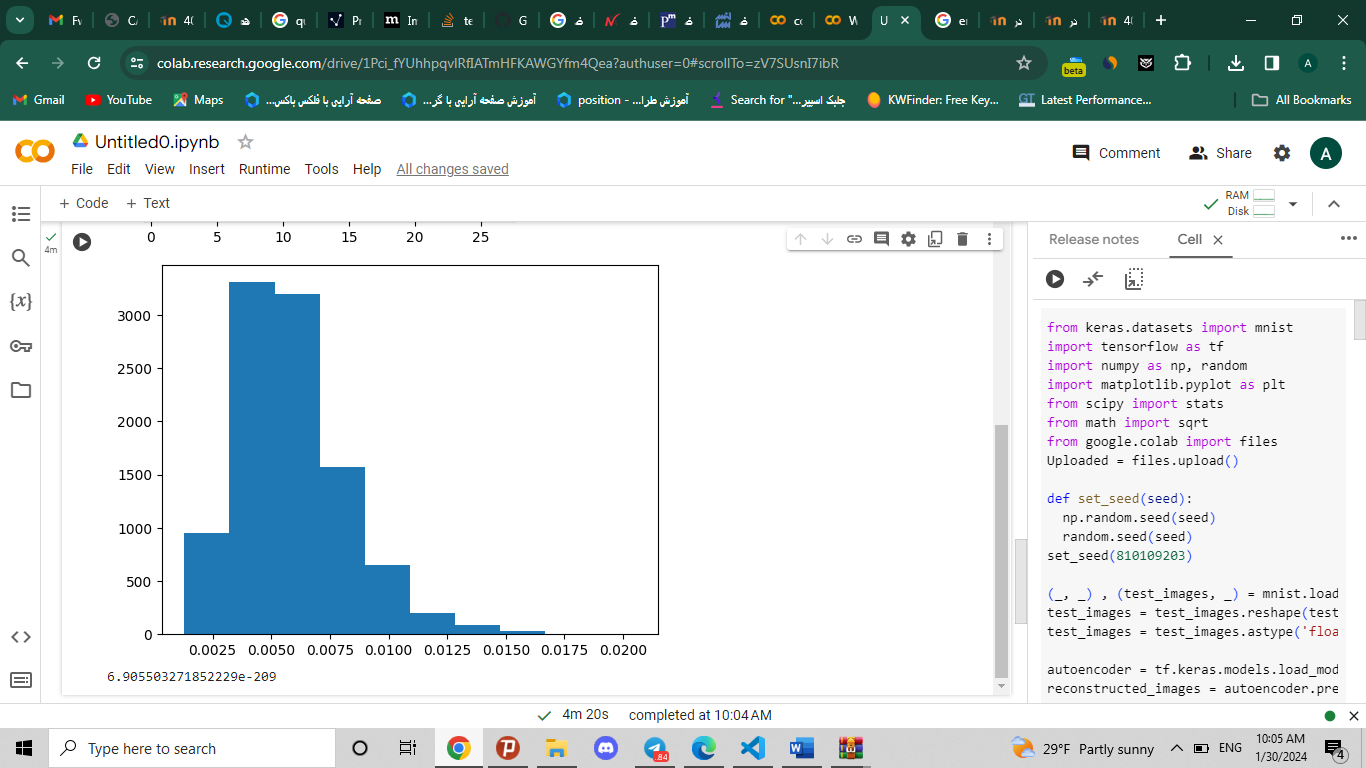






۳ـ د‌ :

کد این قسمت را نیز با توجه به احتلاف داده ها و مربع این اختلاف (با توجه به فرمول اصلی) به دست اوردیم.



که همانگونه که مشخص است مقدار p\_value در پایین histogram نوشته شده و از انجایی که مقدار ان از 0.05 کمتر است نمودار ما از توزیع نرمال پیروی نمی کند. البته در شکل نیز مشخص است که توزیع نرمالی وجود ندار و شکل قرینگی ندارد. البته به دلیل تصادفی انتخاب کردن ۱۰۰۰ نمونه انتظار داریم در هر بار اجرا مقدارp\_value متفاوت باشد.

۳ ـ ه :

براي این قسمت، ابتدا با استفاده از تابع MSE، MSE هر کدام از عکسها را به دست می آوریم و سپس به صورت تصادفی، 1000 نمونه از این مقادیر را انتخاب کرده و میانگین و واریانس این نمونه ها را حساب می کنیم میانگین و انحراف معیار به دست امده را به تابع مربوطه خواهیم داد تا مقدار p\_value را برای ما محاسبه کند. مشاهده می کنیم که مقدار ان از 0.05 کمتر می باشد و می توان نتیجه گرفت که از توزیع نرمال پیروی نمی کند.از انجایی که نمونه ۱۰۰۰ تایی ما به صورت رندوم انتخاب شده است انتظار داریم که مقدار p\_value در هر بار اجرای برنامه متفاوت باشد ولی همچنان در هر بار مقدار ان از 0.05 کمتر خواهد بود که فرقی در نتیجه نهایی نخواهد کرد.

سوال دوم :

۱ـ

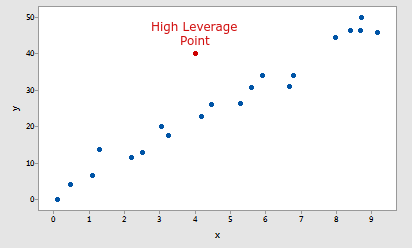
داده پرت (outlier) در [آمار کاربردی](https://parsmodir.com/statistics/statistcs.php) به داده یا داده‌هایی گفته می‌شود که مقدار آن به‌صورت قابل اعتنایی با دیگر داده‌ها و مشاهدات متفاوت باشد. در اصل Outlier مقادیری است که نسبت به میانگین کل داده‌ها فاصله زیادی دارد.

دلایل ایجاد داده‌های پرت :

خطای ابزار، پژوهشگر و پاسخگو مهم‌ترین دلایل ایجاد داده‌های پرت هستند:

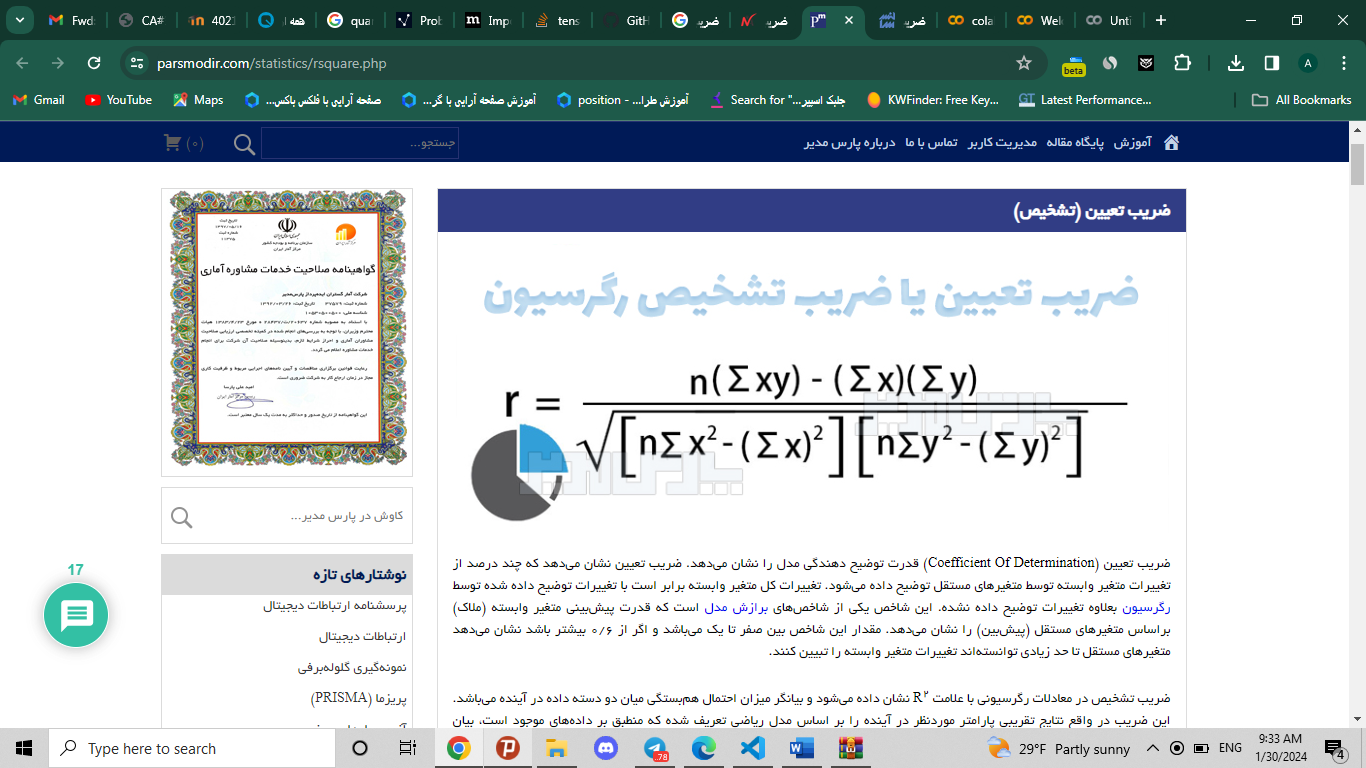
* اشتباه در طراحی مقیاس اندازه گیری (خطای اندازه گیری یا خطای ابزار)
* اشتباه در وارد کردن داده‌ها به نرم افزار (خطای پژوهشگر)
* اشتباه خود پاسخ دهنده به سوالات

نقاط اهرمی یک نقطه داده است که دارای مقدار پیش بینی کننده شدید باشد . یک مقدار ورودی شدید می تواند در مقایسه با سایر نقاط داده در کل مجموعه بسیار کم یه بسیار زیاد باشد. به عنوان مثال در میان داده هاي داده شده، مقدار مقدار x، در بین نقاط با y مشابه، یک جهش ناگهانی داشته و درنتیجه این نقطه، یک نقطه اهرمی است. نقاط اهرمی می توانند در خطی که از نتیجه رگرسیون خطی حاصل می شود نتیجه منفی داشته باشد و مقدار خطای ما را بیشتر کند (همانند نقاط پرت). که در شکل زیر نمونه ای از این نقطه در مقایسه با سایر نقاط خواهیم دید:



به طور خلاصه نقطه نقطه پرت نقطه داده ای است که پاسخ y از روند کلی بقیه داده های پیروی نمی کند در حالی که اگر یک نقطه داده دارای مقادیر x پیش بینی کننده افراطی باشد اهرمی می باشد.

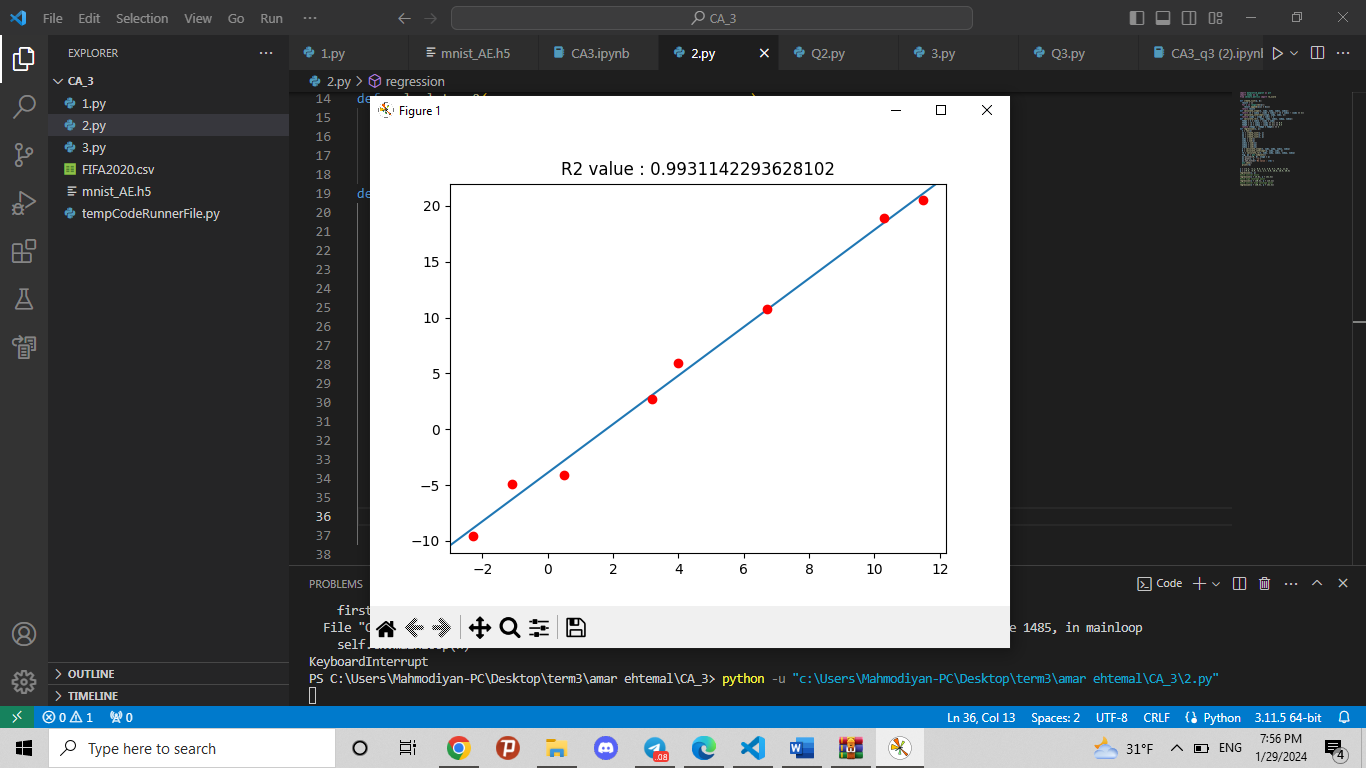
۲ـضریب تعیین در معادلات رگرسیونی با علامت R2 نشان داده می‌شود و بیانگر میزان احتمال هم‌بستگی میان دو دسته داده در آینده می‌باشد. این ضریب در واقع نتایج تقریبی پارامتر موردنظر در آینده را بر اساس مدل ریاضی تعریف شده که منطبق بر داده‌های موجود است، بیان می‌دارد. در بحث رگرسیون خطی، ضریب تعیین را می توان به سادگی برابر با مربع ضریب همبستگی بین مقادیر مشاهده شده و مقادیر پیشبینی کننده دانست و طبق تعریف داریم :



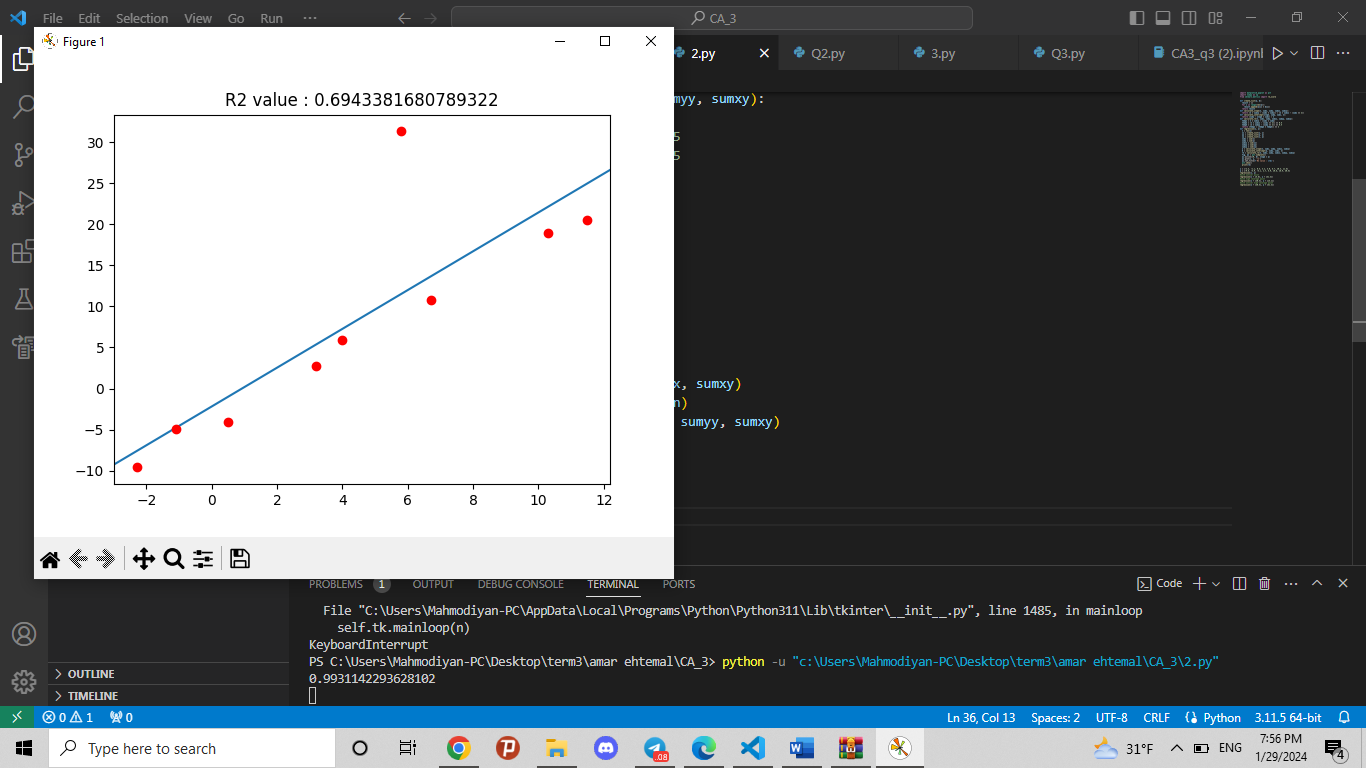
ضریب تعیین در واقع یک ضریبی بین ۰ و ۱ می باشد که بیانگر میزان درستی پیش بینی ما است. هر چه قدر ضریب تعیین به ۱ نزدیک تر باشد نشان دهنده ی این است که پیش بینی ما دقیق تر بوده است که اگر این میزان برابر با یک باشد یعنی پیش بینی ما کامل و بدون نقص بوده است.البته این ضریب را نیز می توان از روی نمودار هایی که رسم کردیم نیز متوجه شد به این صورت که هر چقدر خط حاصل از رگرسیون خطی انطباق و نزدیکی بیشتری به نقاط داده داشته باشد انتضار داریم که ضریب تعیین به ۱ نزدیک تر باشد و همانطور که در بخش های قبلی توضیح دادیم نقاط پرت و اهرمی می توانند در ضریب تعیین ما تاثیر منفی داشته باشند که در شکل های بعدی تاثیرشان را نشان خواهیم داد.

۳ـ

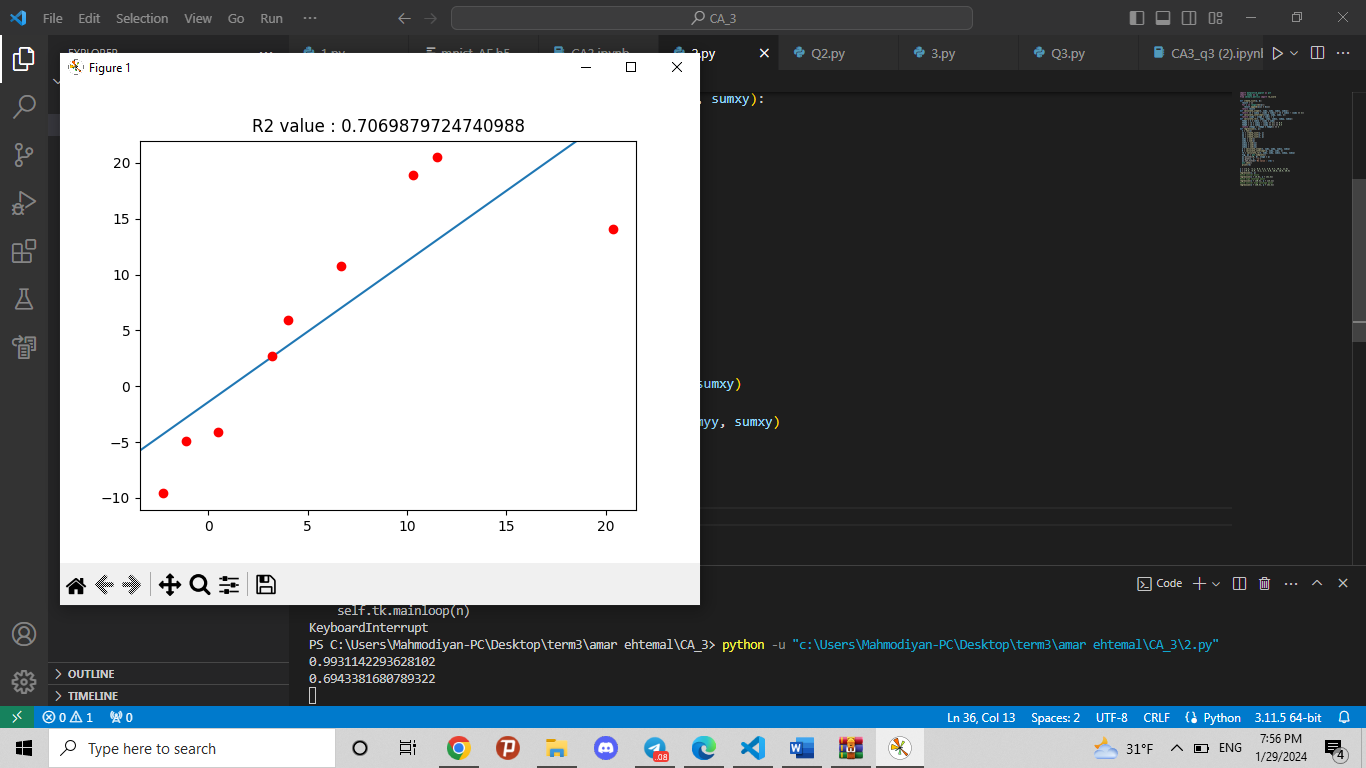
ابتدا نموداری را مشاهده می کنیم که حاصل رگرسیون خطی نقاط عادی ما هستند که شامل ۸ نقطه اولیه هستند که از انجایی که داده ها شمال نقاط پرت و اهرمی نیستند همانگونه که انتظار داشتیم تقریب ما با ضریب تعینن ۹۹درصد انطباق بسیار بالایی با نقاط داده دارد.



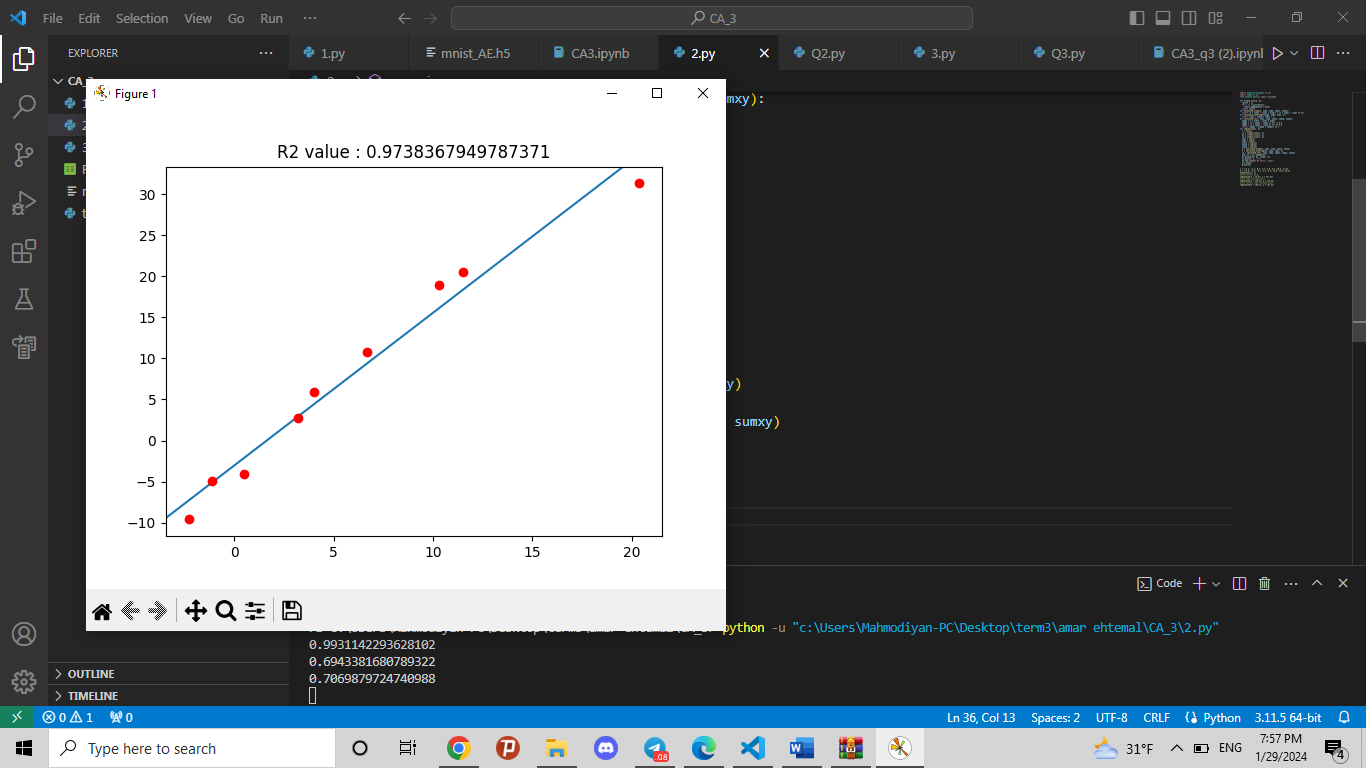
سپس نقطه ی دور افتاده را به ۸ نقطه ی اصلی اضافه می کنیم که در شکل نیز این نقطه مشخص است که روند متفاوتی با سایر نقاط دارد که طبق انتظار نیز تاثییر منفی خود را روی ضریب تعیین گذاشته و ان را تا مقدار ۶۹ درصد کاهش داده.



در شکل بعدی نقطه ی اهرمی را به ۸ داده اصلی اضافه می کنیم و دوباره طبق انتظار این نقطه دوباره تاثیر منفی رو ضریب تعیین گذاشته ولی تاثیر منفی ان از نقطه ی پرت کمتر بوده و مقدار صریب تعیین را تا ۷۰ درصد کاهش داده.



در نمودار آخر، نقطه ی دورافتاده ـ اهرمی را شامل داده های خود کردیم و می تونایم مشاهده کنیم که این نقطه تفریبا روند یکسانی با سایر نقاط دارد و باعث انحراف کمی در خط حاصل از رگرسیون خطی شده و همچنان ضریب تعیین مقدار خوبی داشته و بسیار به یک نزدیک است که مقدار ان در بالای تصویر قابل مشاهده است.



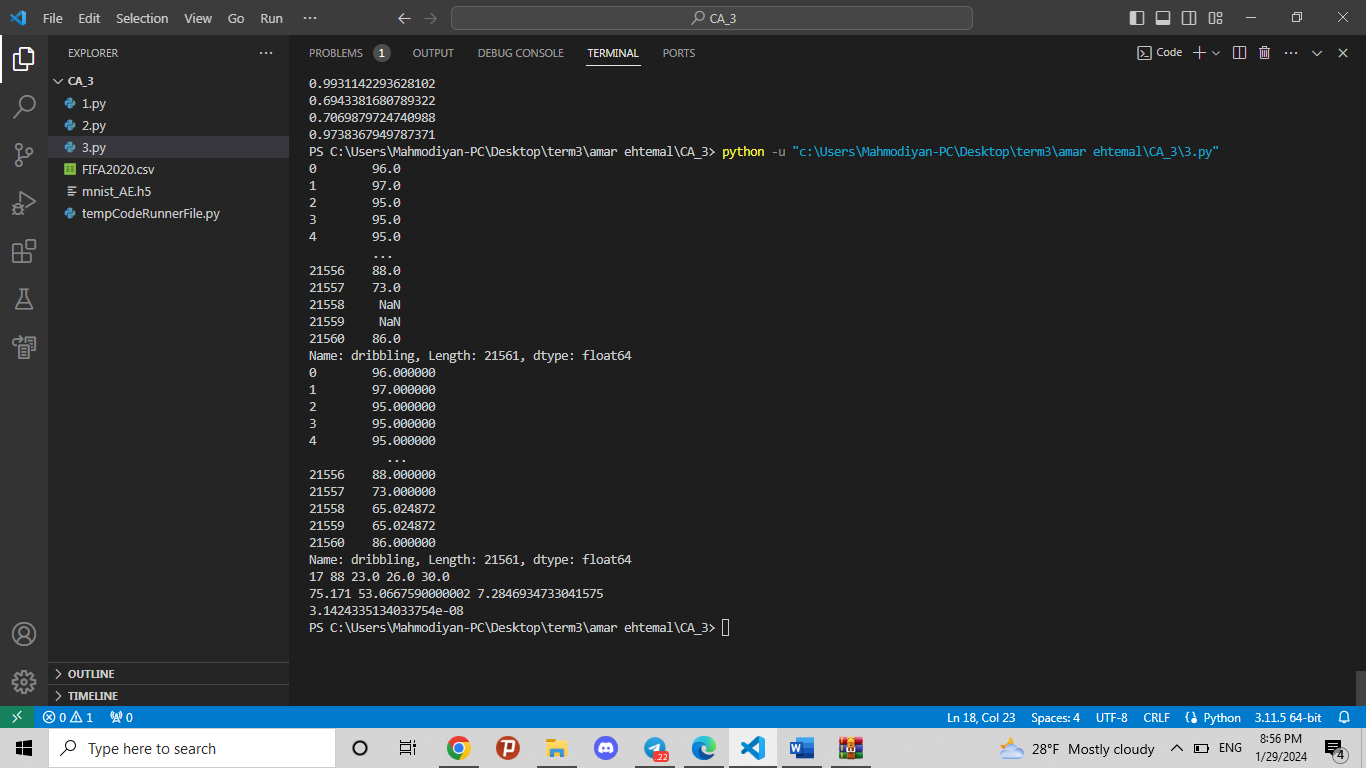
۴‌ـ

همانطور که دیدیم روش کمترین مربع خطا حساسیت زیادی به داده های پرت دارد که یکی از روش های مقابله با ان این است که قبل از مدل کردن داده بیاییم و داده های دور افتاده را از بین داده های خود حذف کنیم تا مدل بهتری داشته باشیم گرچه بهتر است فقط زمانی اینکار را انجام دهیم که دلیل موجهی برای ان داریم زیرا برخی نقاط حتی با وجود دور افتادگی نشان دهنده ی تغییرات طبیعی در جمعیت هستند و باید همانطور که هستند در مجموعه داده های ما باقی بمانند. همچنین حذف این نقاط می تواند باعث کوچک تر شدن نمونه و یا از دست رفتن اطلاعات مهم شود.

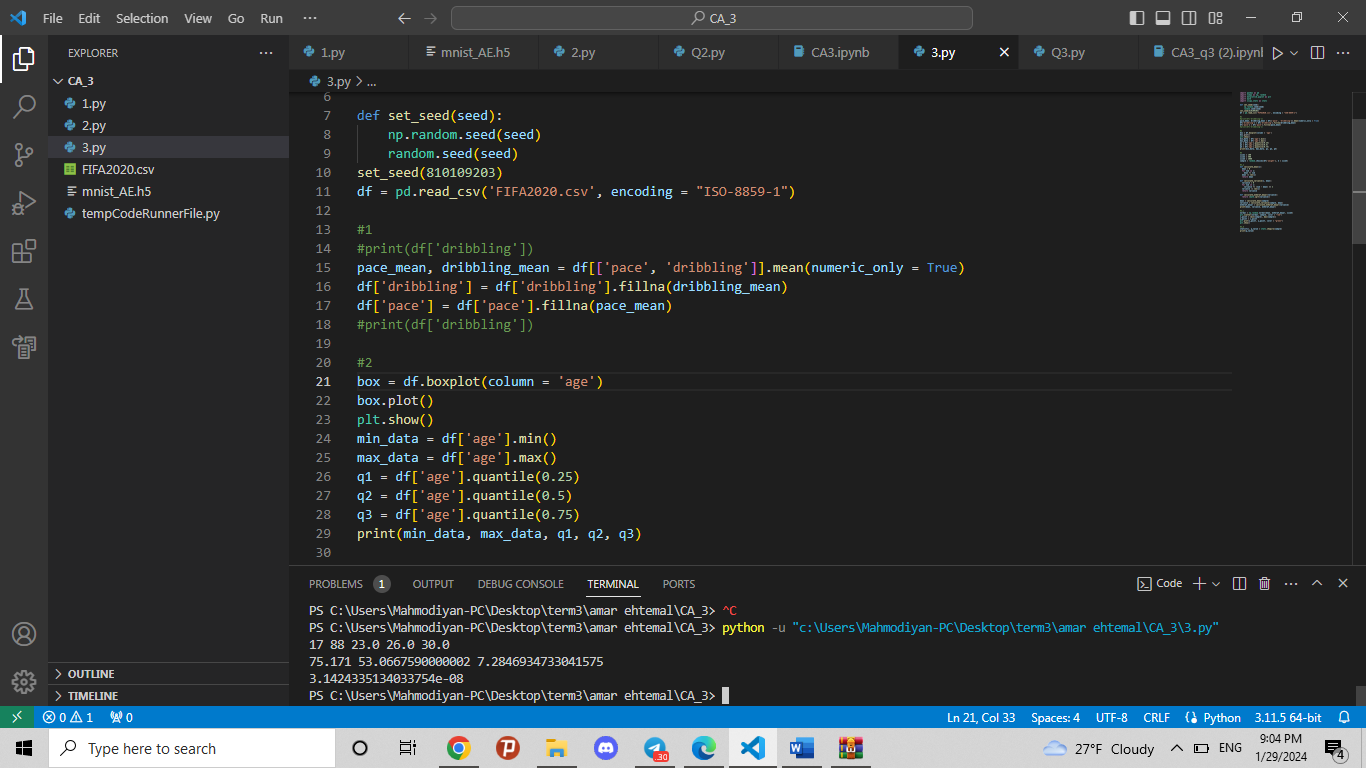
البته روش های دیگری برای مقابله با این نقاط هستند در واقع باید از روش هایی استفاده کنیم که حساسیت کمتری نسبت به این نقاط داشته باشند. برای افزایش مدل رگرسیون خطی در برابر نقاط \رت استفاده از تکنیک های رگرسیون قوی مانند رگرسیون هوبر یا RANSAC را در نظر می گیریم که کمتر تحت تاثیر نقاط داده شدید قرار می گیرند. علاوه بر ان استفاده از مقیاس بندی ویژگی و نرمال سازی می تواند تاثیر نقاط پرت را کاهش دهد.

سوال سوم :

۱‌- بهترین راه این است که این مقادیر را با مقدار میانگنین ستون مربوطه شان جایگزین کنیم چرا که این مقادیر در میانگین نهایی بی تاثیر هستند و ماینگین داده های ما تغییری نخواهد کرد پس برای اینکار ابتدا برای ستون های pace وdribbling مقدار های N/A را به NAN تبدیل می کنیم سپس میانگین ستون مزبوطه را گرفته و توسط تابه fillna این مقدار را با مقدار NAN جایگزین می کنیم .که در شکل زیر این مساله قابل رویت است :



۲ـ مقادیر خواسته شده را به دست اورده و چاپ می کنیم :



که این مقادیر به ترتیب نشان گر‌:

Max\_age = 88

Min\_age = 17

Q1 = 23

Q2 = 26

Q3 = 30

که همانگونه که از چارک ها قابل مشاهده است سن بیشتر افراد در بازه ی بین ۲۳ تا ۳۰ سال می باشد و می توان نتیجه گرفت که سن ۸۸ یک داده پرت در مقایسه با دیگر داده ها به حساب می اید.

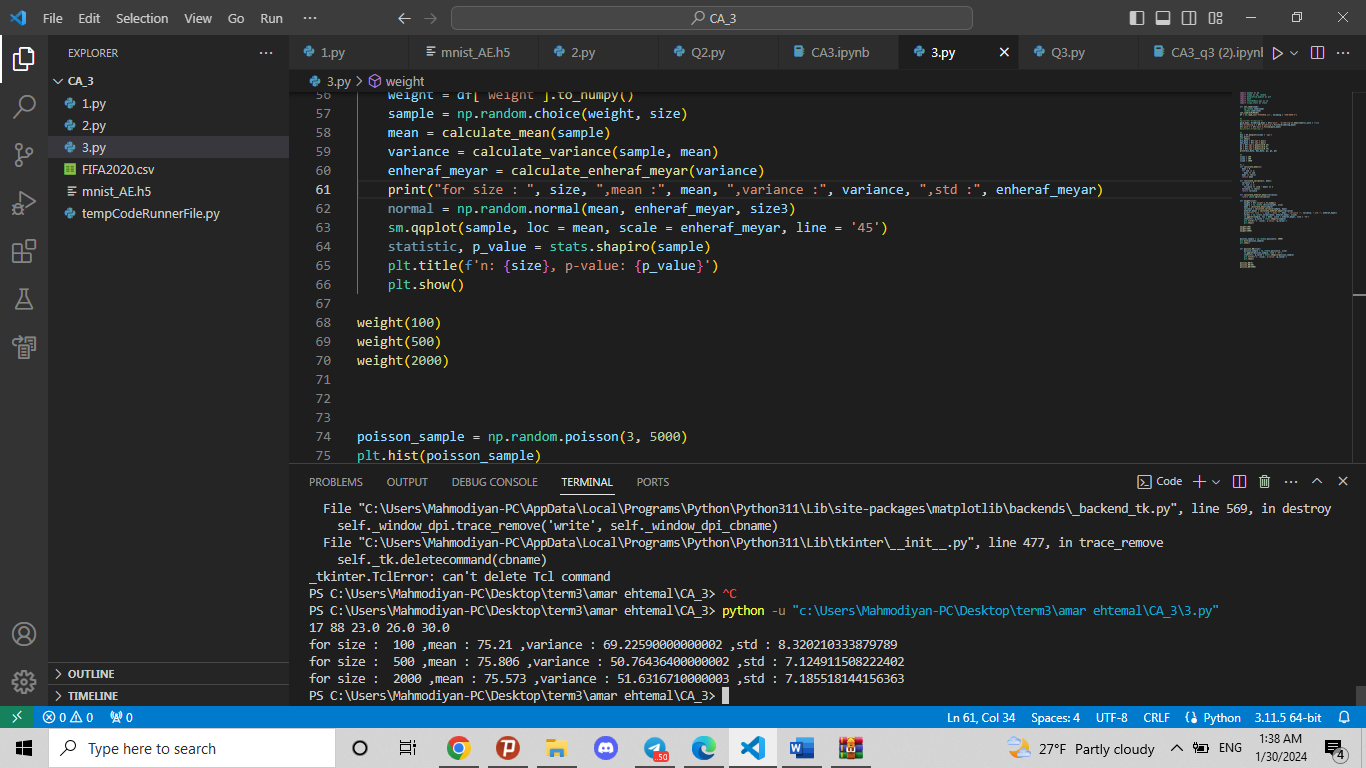
Q1 : مشاهده ای از مجموعه داده های مورد بررسی است که یک چهارم داده ها (یعنی ۲۵ درصد مشاهدات) از آن کوچکتر و سه چهارم داده ها (یعنی ۷۵ درصد مشاهدات) از آن بزرگتر می باشد.

Q2 : چارک دوم همان میانه می باشد، داده ای که ۵۰ درصد (نیمی) از مشاهدات از آن کوچکتر یا مساوی و ۵۰ درصد (نصف دیگر) از آن بزرگتر می باشند.

Q3 : مشاهده ای از مجموعه داده های مورد بررسی است که سه چهارم داده ها (یعنی ۷۵ درصد مشاهدات) از آن کوچکتر و یک چهارم داده ها (یعنی ۲۵ درصد مشاهدات) از آن بزرگتر می باشد.

۳ـ الف ـ

میانگین و واریانس و انحراف معیار را برای نمونه های ۱۰۰ تایی ۵۰۰ تایی و ۲۰۰۰ تایی به دست اورده :



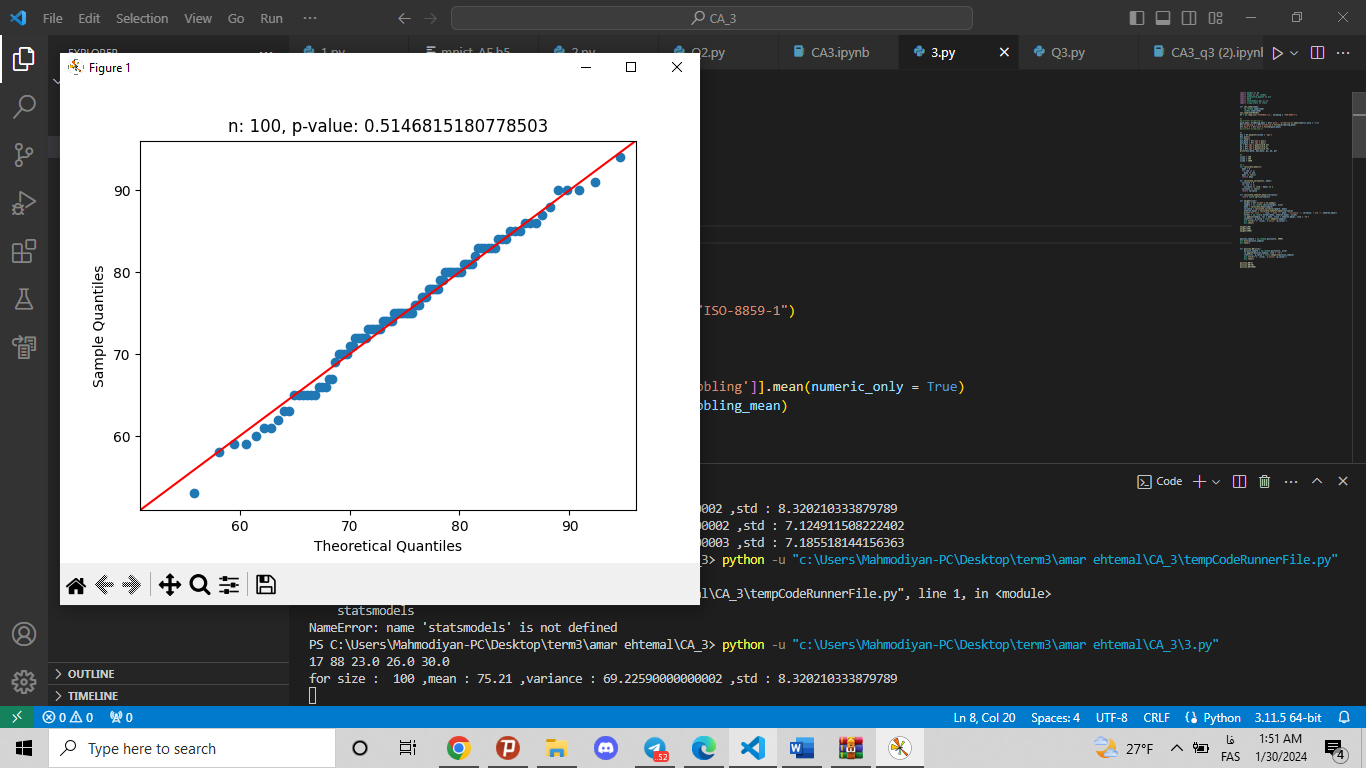
۳ـ ب :

نمودار چندک چندک یا Q-Q plot به منظور مقایسه دو توزیع به کار گرفته می‌شود. از چنین نمودارهایی حتی می‌توان مطابقت توزیع داده‌ها را با یک توزیع مشخص، مورد بررسی قرار داد. توسط نمودار چندک چندک یا Q-Q plot شکل توزیع‌ها مقایسه می‌شود و یک تصویر گرافیکی یا نمودار برای نمایش میزان مطابقت آن دو توزیع نشان داده می‌شود. چولگی، پارامتر مرکزی و پراکندگی در نحوه مقایسه دو توزیع مشکلی ایجاد نمی‌کنند و به راحتی می‌توان هم توزیع (Equal Distributed) بودن داده‌ها را دو گروه مقایسه کرد.

همچنین نمودار چندک-چندک (Quantile-Quantile Plot) یک ابزار توصیفی است که کمک می‌کند بفهمیم یک مجموعه داده از یک توزیع احتمالی مشخص مانند توزیع نرمال (Normal Distribution) یا نمایی (Exponential Distribution) پیروی می‌کند یا خیر. برای مثال فرض نرمال بودن داده‌ها در بسیاری از آزمون‌های آماری وجود دارد. بنابراین قبل از انجام چنین آزمون‌های آماری نیاز است تا بررسی کنیم که داده‌ها از توزیع نرمال پیروی می‌کند. یک ابزار ساده و قدرتمند در چنین تحلیلی، استفاده از نمودار چندک-چندک است. اگر نمونه داده‌ها از توزیع نرمال پیروی کند، باید داده‌ها نزدیک به خط ۴۵ درجه باشند.

۳- ج :

نمودار انرا با کمک کتابخانه statsmodel رسم می کنیم و می بینیم به ازای n = 100 نقاط داده تقریبا انطباق خوبی با خط y = x دارند. یعنی توزیع ما به توزیع نرمال نزدیک است. همچنین انتظار داریم که هرچه n ما بزرگتر باشد طبق قضیه حد مرکزی توزیه ما نرمال تر شود و بر خط نیم ساز منطبق تر شود که در بخش ه ) مشاهده این امر برای n های بزرگتر خواهیم بود.

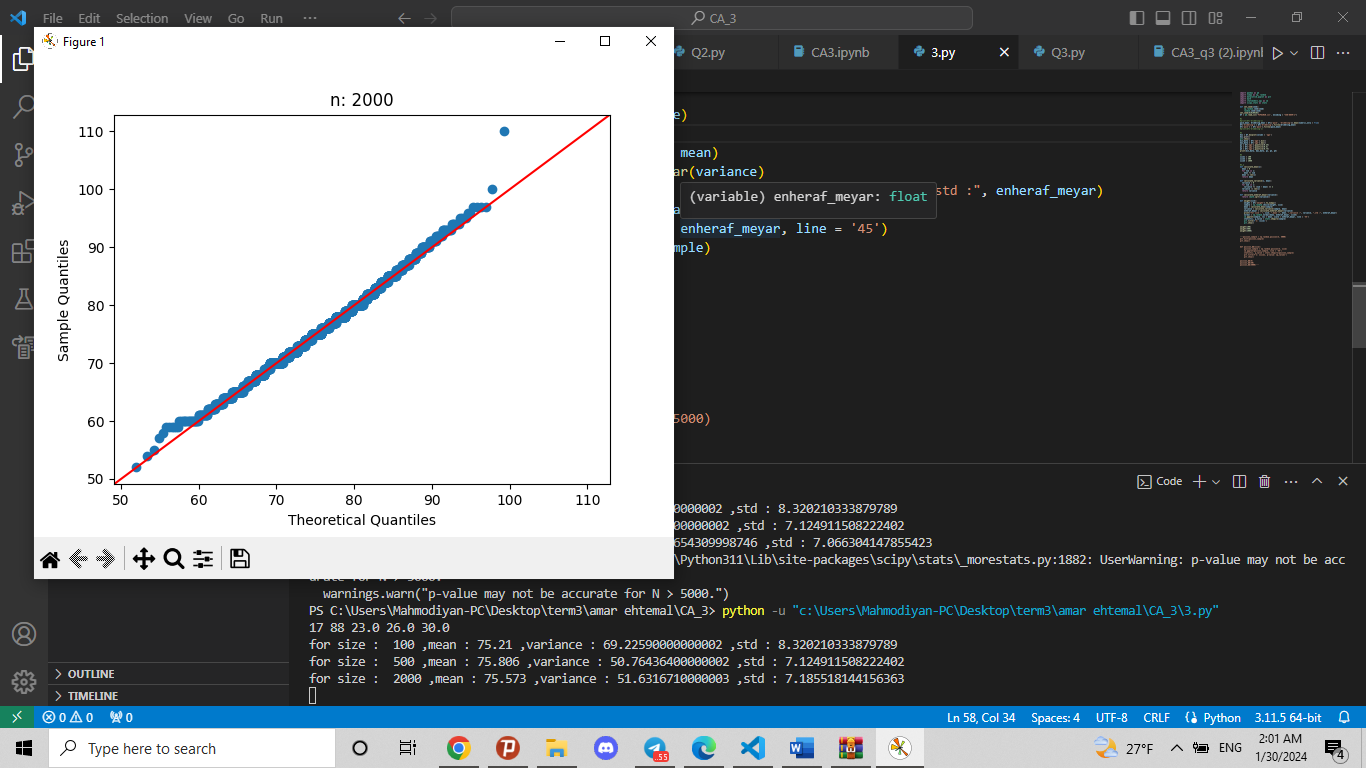
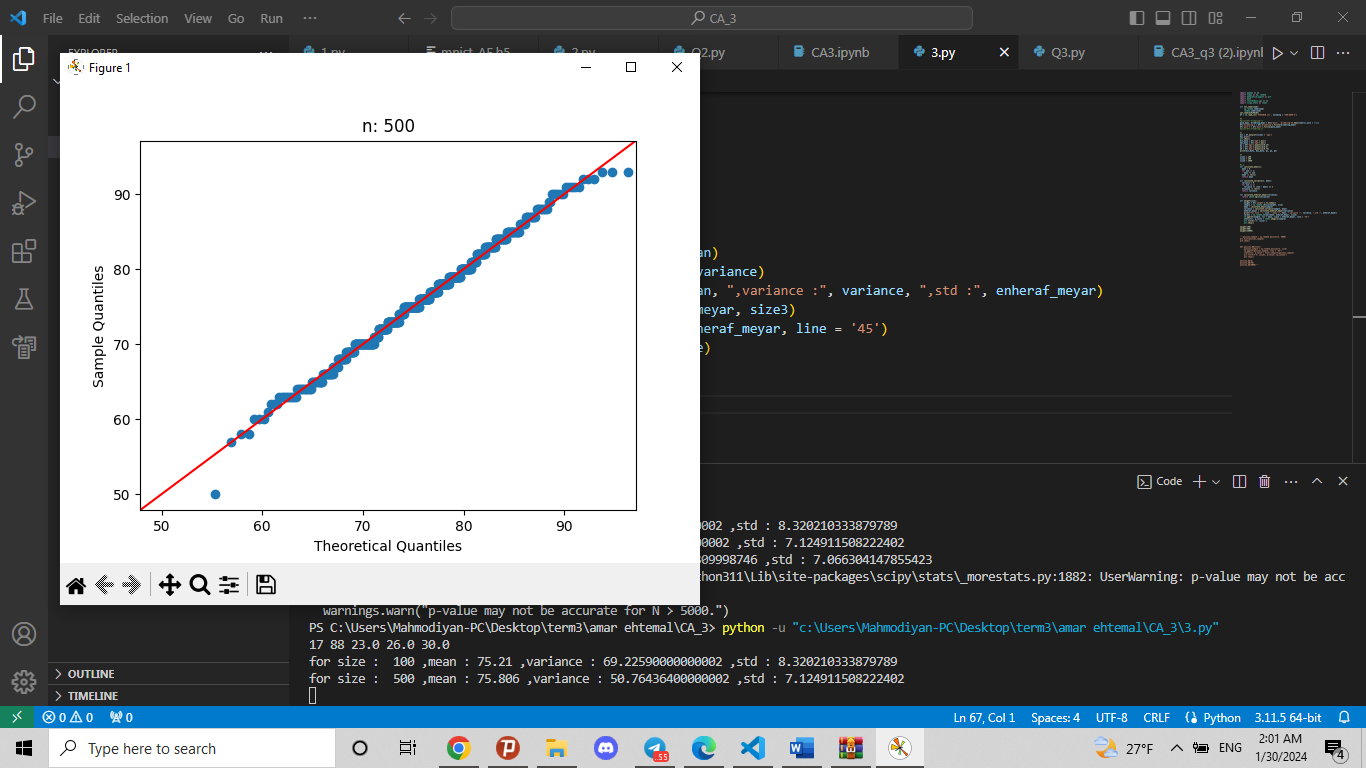


۳ ـ د :

انتظار داریم با هر بار انجام برنامه با توجه به تصادفی انتخاب شدن نمونه ها مقدار p\_value در هربار متفاوت باشد ولی در این قسمت مقدار ان برابر 0.51 بوده که بزرگتر از 0.05 است در نتیجه توزیع را یک توزیع نرمال در نظر می گیریم.

۳ ـ ه :

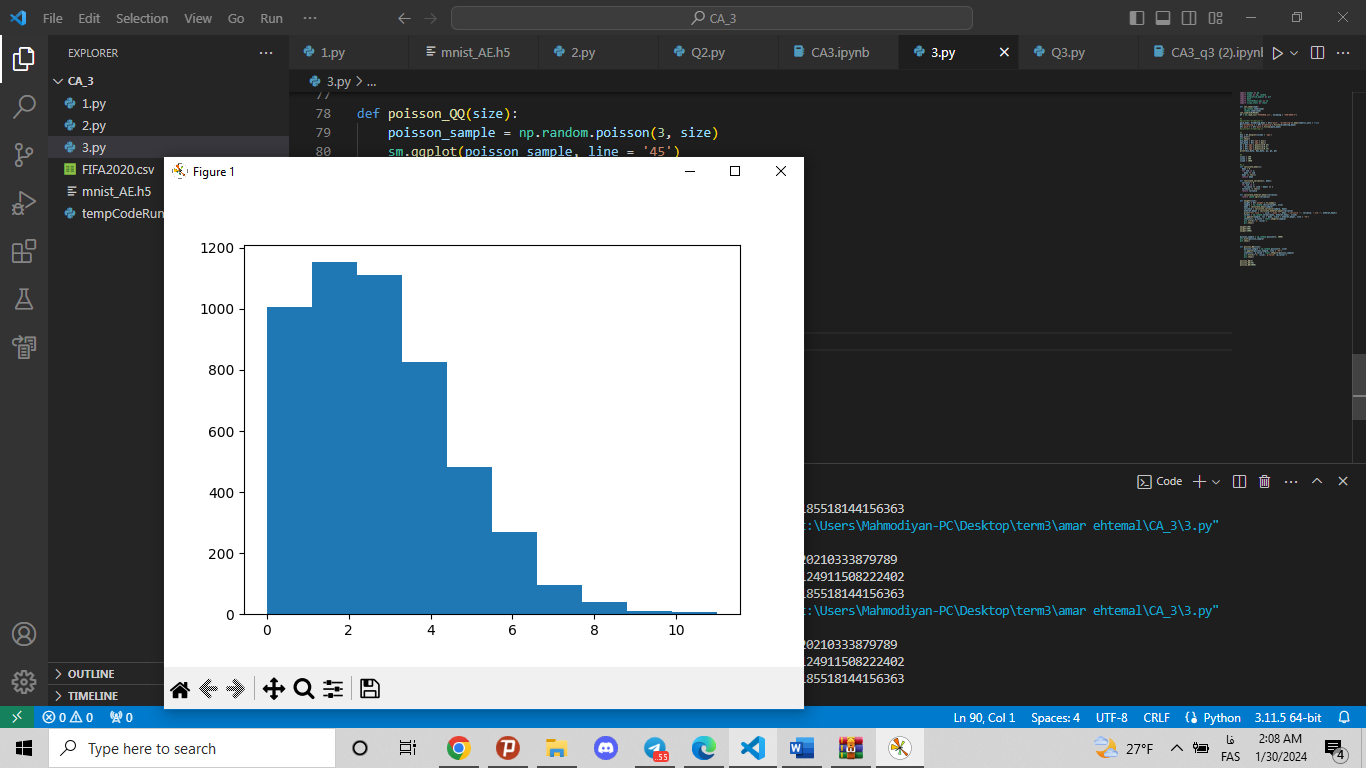
مراحل قبل را برای n = 500 و n = 2000 انجام می دهیم که نتیجه در عکس های زیر قابل مشاهده است:



همانگونه که مشاهده می شود با افزایش n نقاط داده انطباق بیشتری بر روی نیمساز دارند که یعنی توزیع ما با افزایش n به توزیع نرمال نزدیک تر می شود.

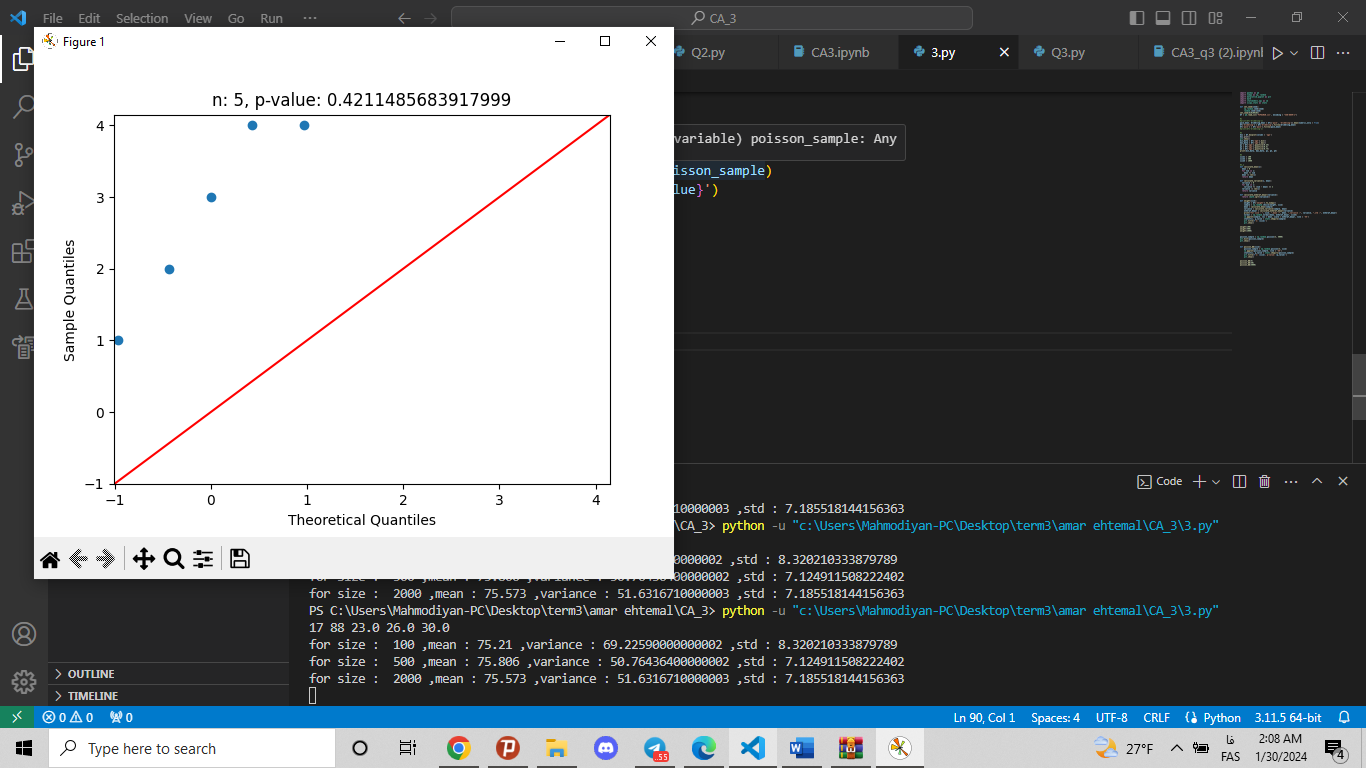
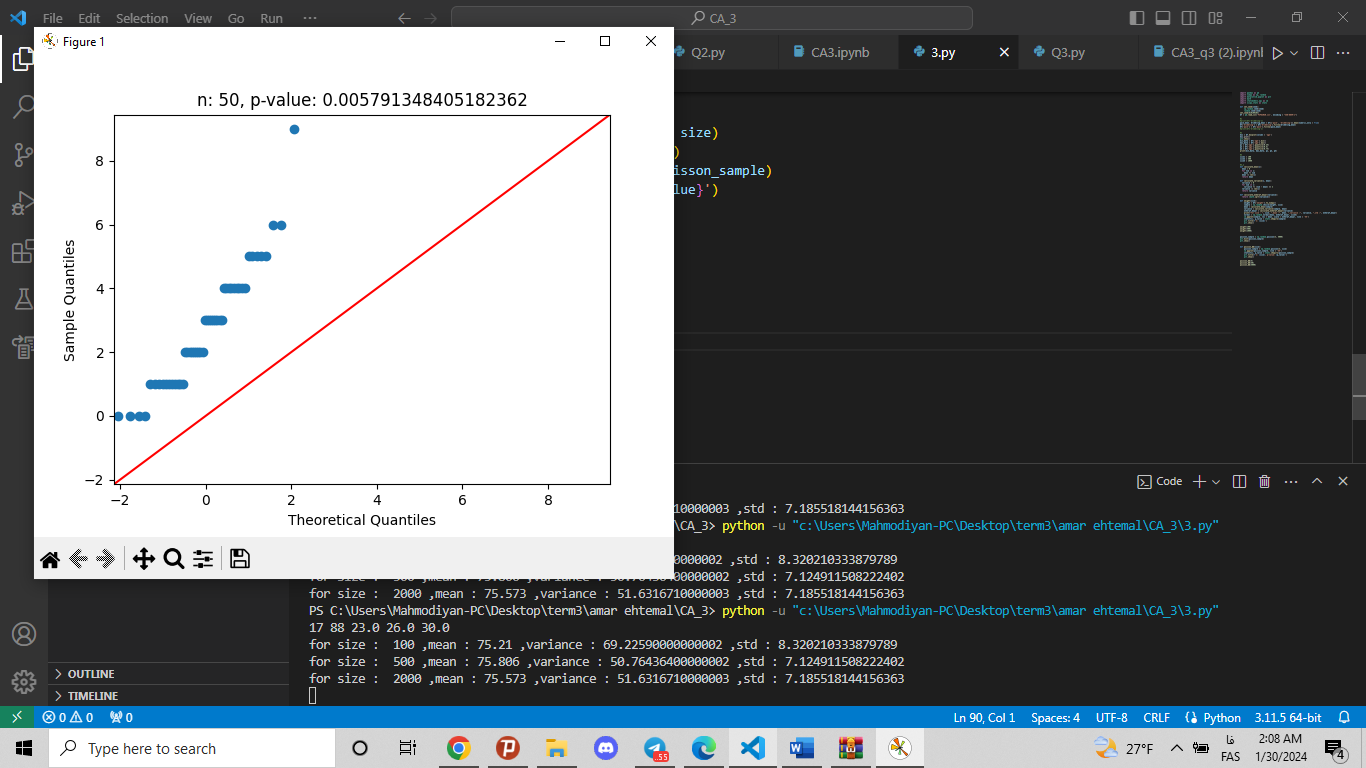
۴ـ الف :

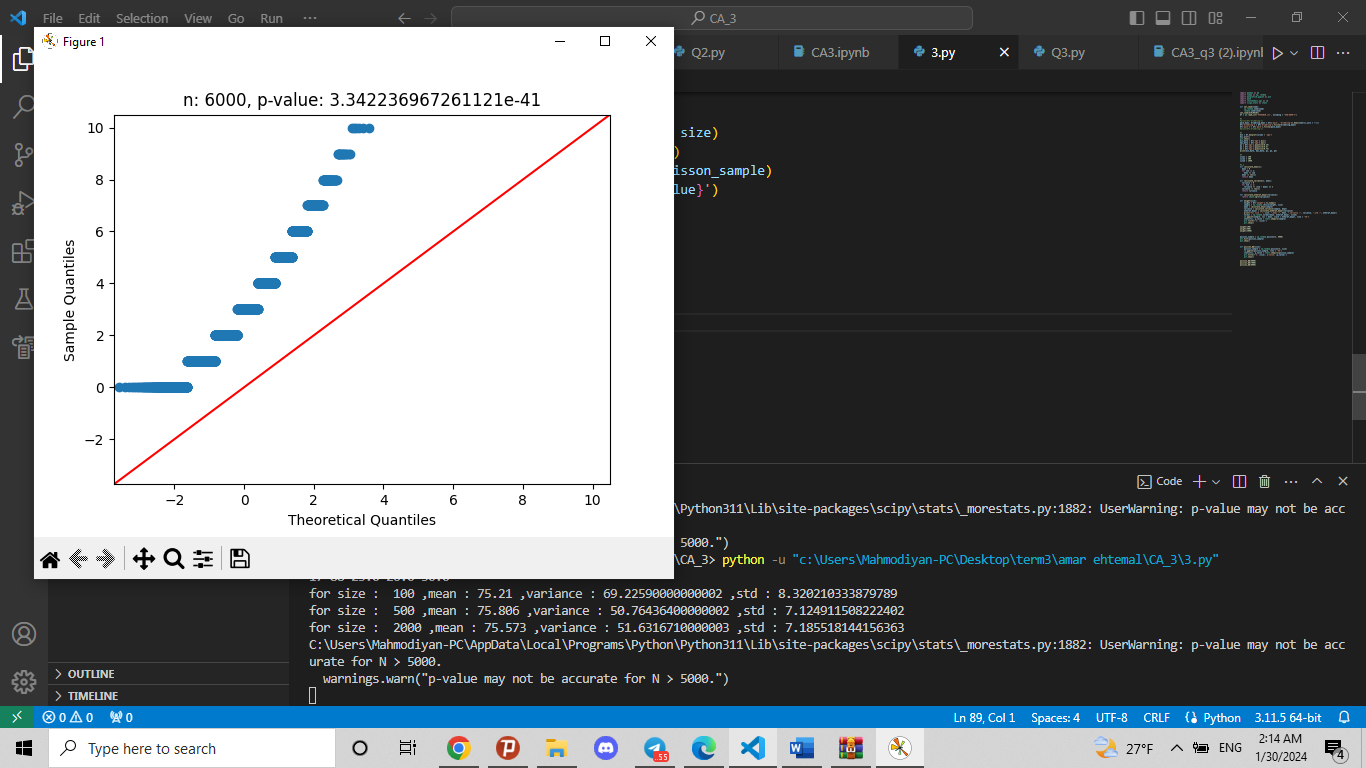
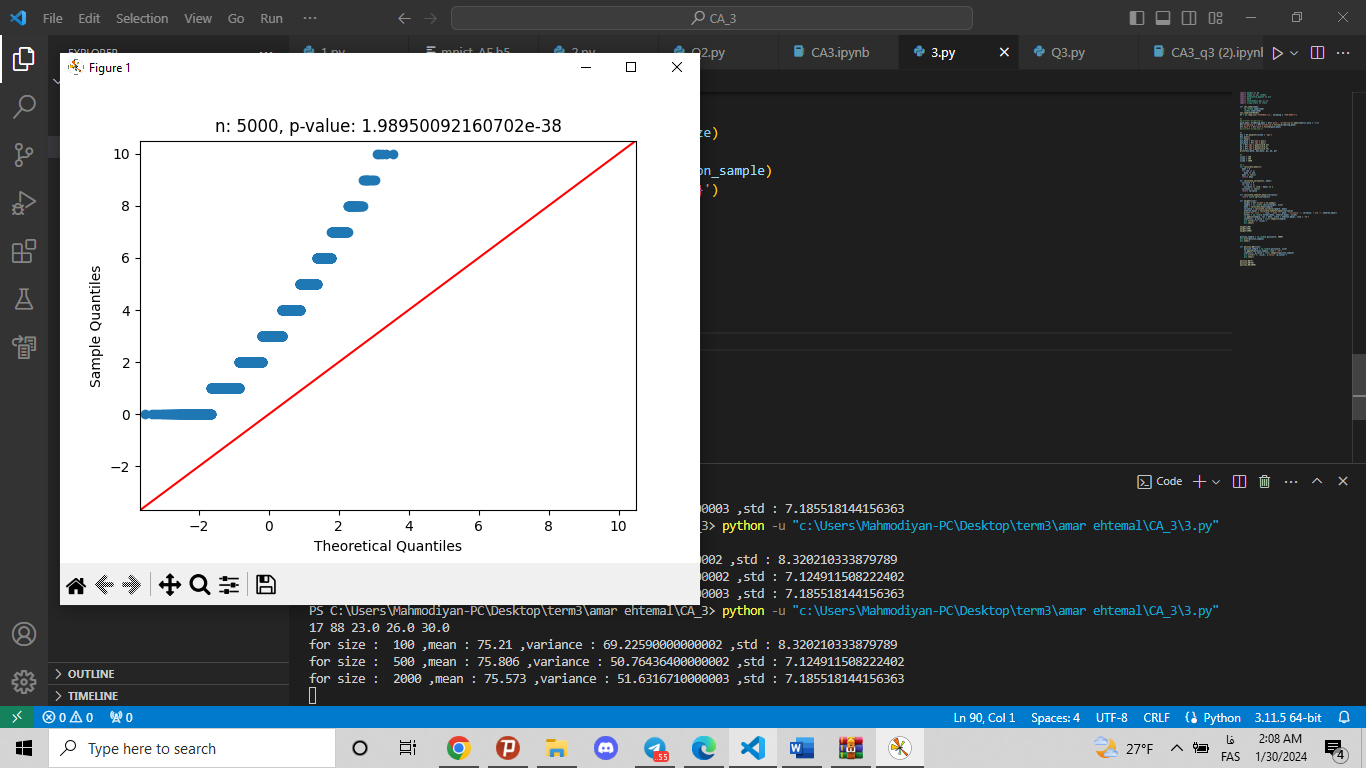
نمودار هیستوگرام :

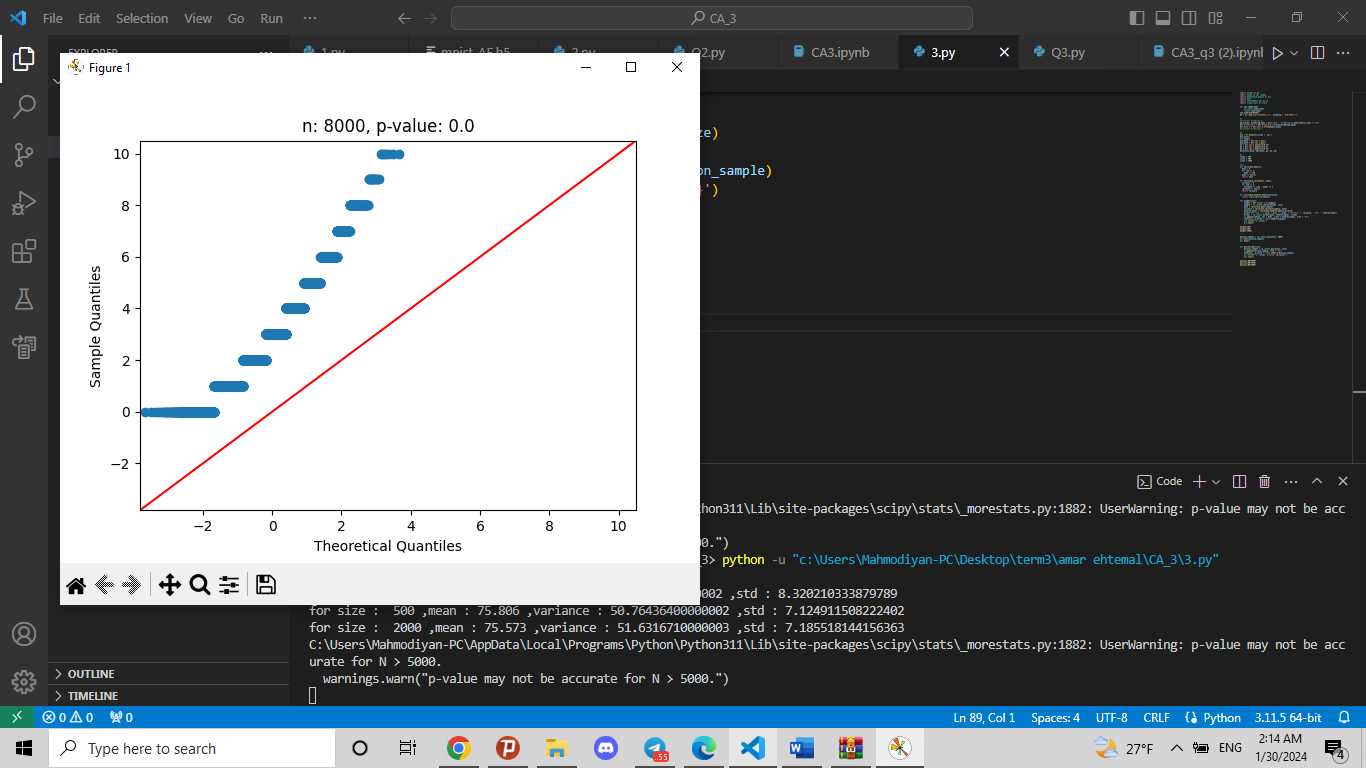
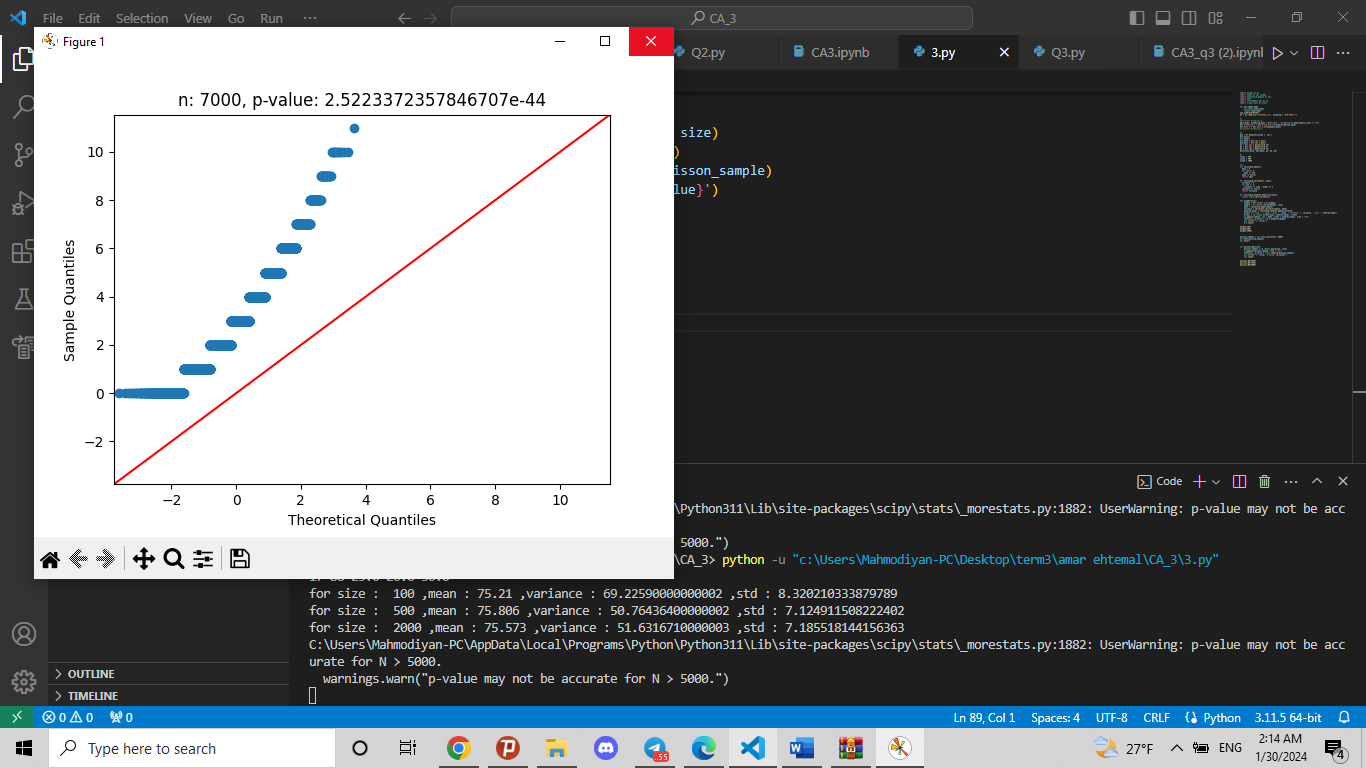


۴ـ ب :

نمودار ها:







از نمودار ها و مقدار p\_value ها می توان نتیجه جالبی گرفت اینکه از انجایی که توزیع پواسون به ازای n های کم هنوز شکال نگرفته و مشاهده می شود که در این مثال به ازای n = 6000 نمودار که به توزیع نرمال نزدیک می باشد اما پس از ان هرچقدر n بیشتر می شود توزیع پواسون بیشتر شکل خود را می گیرد و توزیع از حالت نرمال دورتر می شود که می بینیم درنهایت به ازای n های بیشتر از 8000 مقدار p\_value کامل مساوی صفر می شود و توزیع ما دیگر نرمال نیست و کاملا از توزیع پواسون پیروی می کند. . این مورد ممکن است تا حدودي با انتظار ما از قضیه حد مرکزي تفاوت داشته باشد چرا که داده ها همگی دارای توزیع پواسون هستند و با افزایش تعداد داده های پواسون ، نمودار کلی متفاوت با توزیع نرمال خواهد بود

**پایان گزارش**