سوال 1)

1- دو مورد از کاربرد های اتوانکودر ها:

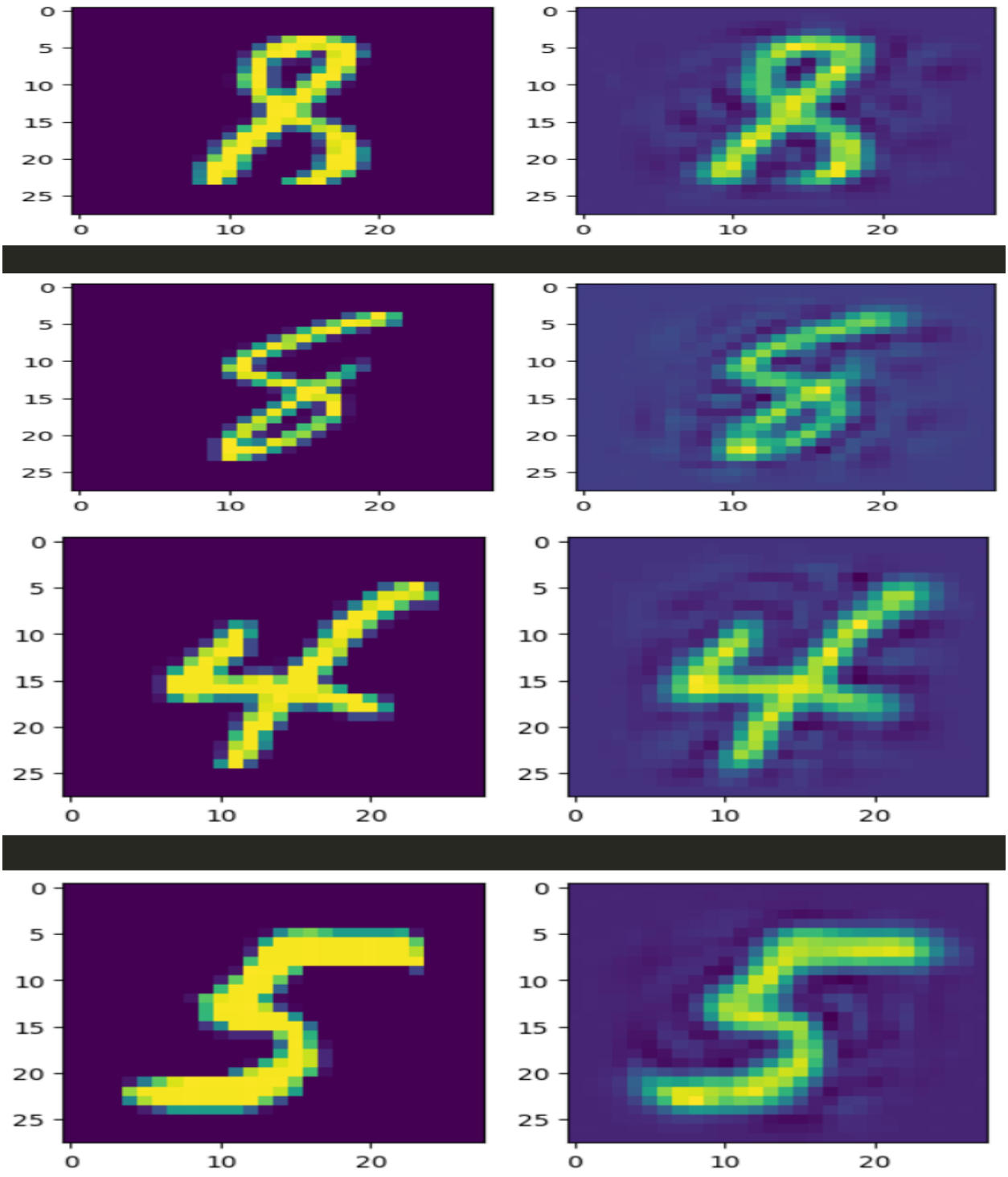
- نویز تصاویر را میتوانند بسیار کاهش دهند یا در مواردی حتی آن ها را از بین ببرند.
- همچنین پس از استفاده از اتوانکودر برای داده ها میتوان ابعاد آن ها را کاهش داد.

2- علت خطای مشاهده شده در اتوانکودر ها اختلاف بین داده های ورودی و خروجی میباشد. میان فضای پنهان یا latent و این خطا و ابعادشان ارتباط تنگاتنگی وجود دارد و روی یکدیگر اثر گذار هستند به طوری که latent نباید خیلی کوچک یا خیلی بزرگ باشد. اگر سایز و ابعاد این فضا نسبت به داده ورودی خیلی کوچکتر باشد اتوانکدور قادر نخواهد بود تمامی اطلاعات مهم داده را انتقال دهد. همچنین اگر ابعاد این فضا نسبت به داده ورودی خیلی بزرگتر باشد نیز هنگام نوشتن داده در ابعاد بزرگتر ممکن است خطا رخ دهد و بدتر شود. در نتیجه از اتوانکودر استفاده میکنیم تا نسبت ابعاد این دو را به خوبی پیدا کنیم و به مشکل نخوریم.

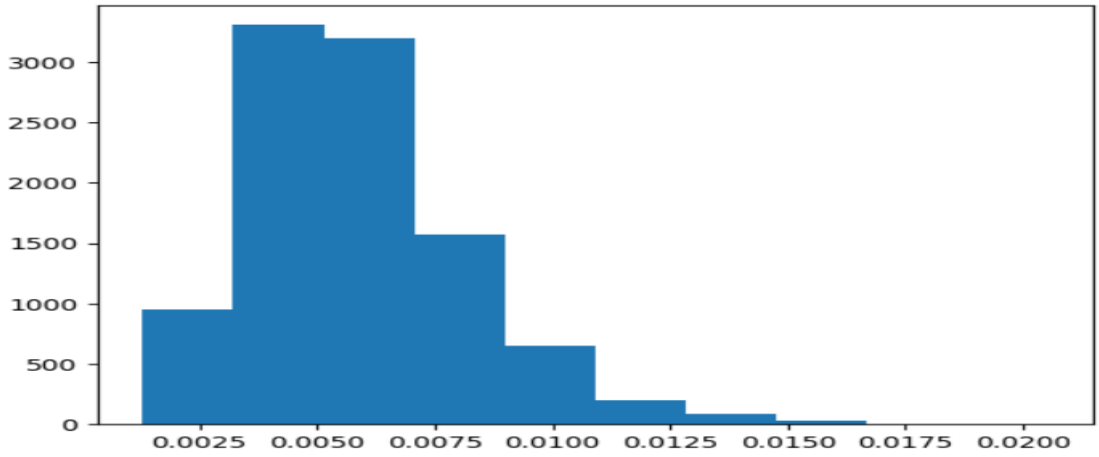
3-

ج) در این قسمت به رسم تصویر اصلی و تصویر انکود شده و مقایسه آن ها پرداختیم. همانطور که مشهود است تصویر reconstructed دارای کیفیت پایین تری است اما مهم این است که تصویر اصلی را به خوبی نمایش میدهد و اطلاعات مهم به درستی انتقال یافته است.

در تصاویر صفحه بعد 4 نمونه تصادفی از تصاویر نشان داده شده است. تصاویر سمت چپ اصلی و تصاویر سمت راست reconstructed آن ها میباشند:



د) پس از محاسبه MSE برای تمامی این تصاویر نمودار هیستوگرام آن ها به شکل زیر میباشد:



ه) در انتها بر اساس آزمون کولموگروف بررسی کردیم که این MSE ها دارای توزیع نرمال میباشند یا خیر. بدین منظور p-value را محاسبه کردیم و طبق آن اگر بیشتر از 5% باشد توزیع نرمال خواهد داشت. نتیجه به این صورت شد:

```
4.802906034637 21e-38
Not Normal
```

همانطور که معلوم است p-value دارای مقدار بسیار کوچکی است و لذا توزیع نرمال نخواهد بود.

سوال 2)

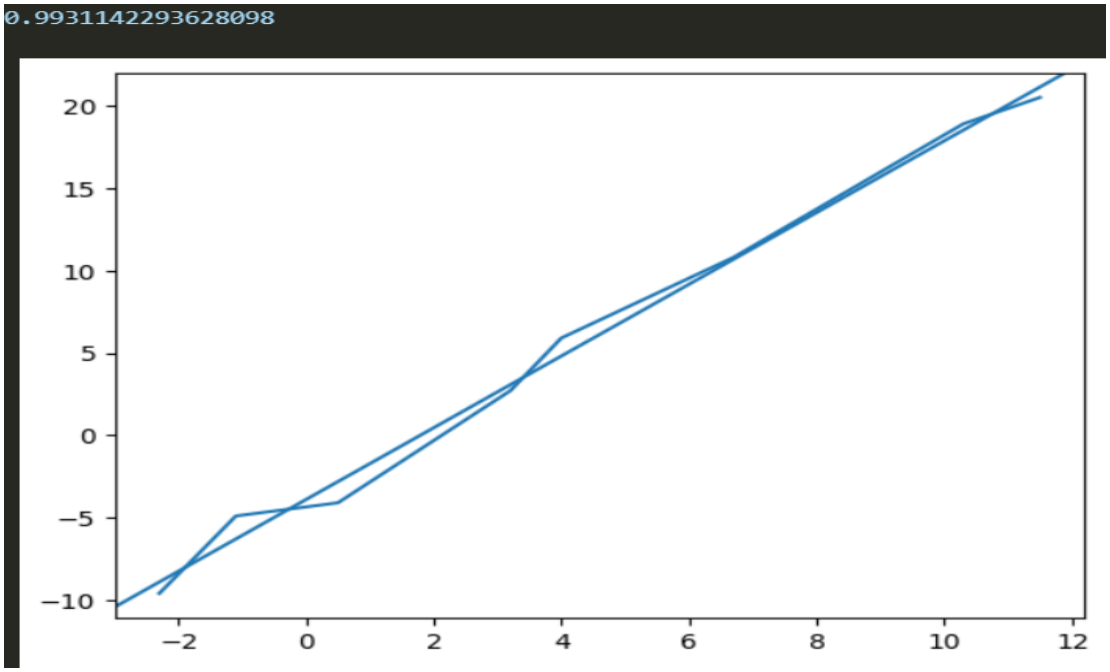
1-

نقاط پرت: این نقاط از بقیه نقاط داده ها فاصله به نسبت زیادی دارند و رفتارشان مانند اکثریت داده ها نیست و اصطلاحا پرت یا دور افتاده هستند. این نوع نقاط ممکن است به علت خطا در جمع آوری داده ها و یا تغییر ناگهانی آن ها بوجود بیایند. باید سعی شود هنگام وجود این نقاط آن ها را از داده های خود به گونه ای حذف کنیم چرا که باعث انحراف زیاد رگرسیون و تغییر شیب آن میشوند.

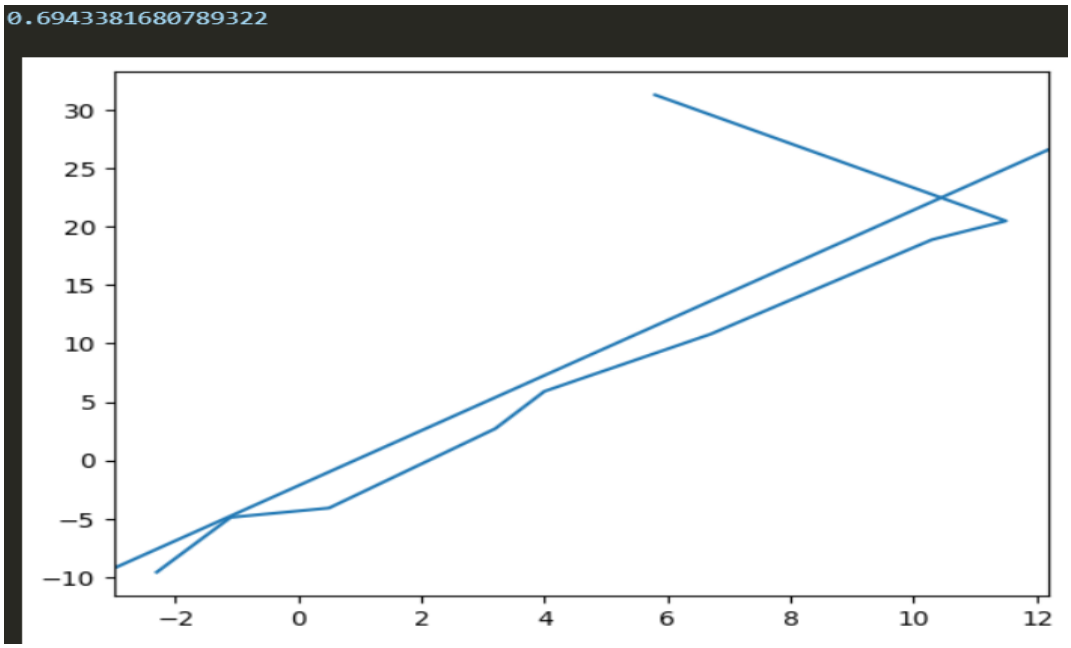
نقاط اهرمی: این نوع نقاط از داده ها نقاطی هستند که مقدارشان نسبت به بقیه نقاط خیلی زیاد یا خیلی کم باشد. یعنی اختلاف آن ها از نقاط مجاور خود زیاد باشد. این نقاط نیز باعث انحراف و غلط شدن نتیجه رگرسیون خواهد شد.

2- ضریب تعیین یا در اینجا همان ضریب همبستگی همواره عددی بین 1- و 1 است و میزان خطی بودن رابطه را اندازه میگیرد. به عبارتی همان کواریانس نرمالیزه است. این ضریب در واقع معیاری برای اندازه گیری میزان درست بودن پیشبینی تولید شده نسبت به تغییرات مختلف در داده خروجی میباشد. به طوری که هرچقدر به 1 یا 1- نزدیکتر باشد یعنی پیشبینی ما (رگرسیون) دقیق تر بوده است و هرچقدر به 0 نزدیکتر باشد یعنی تخمین خوبی نبوده است. در بحث رگرسیون خطی برای فرمول ضریب تعیین از همان فرمول ضریب همبستگی استفاده میکنیم.

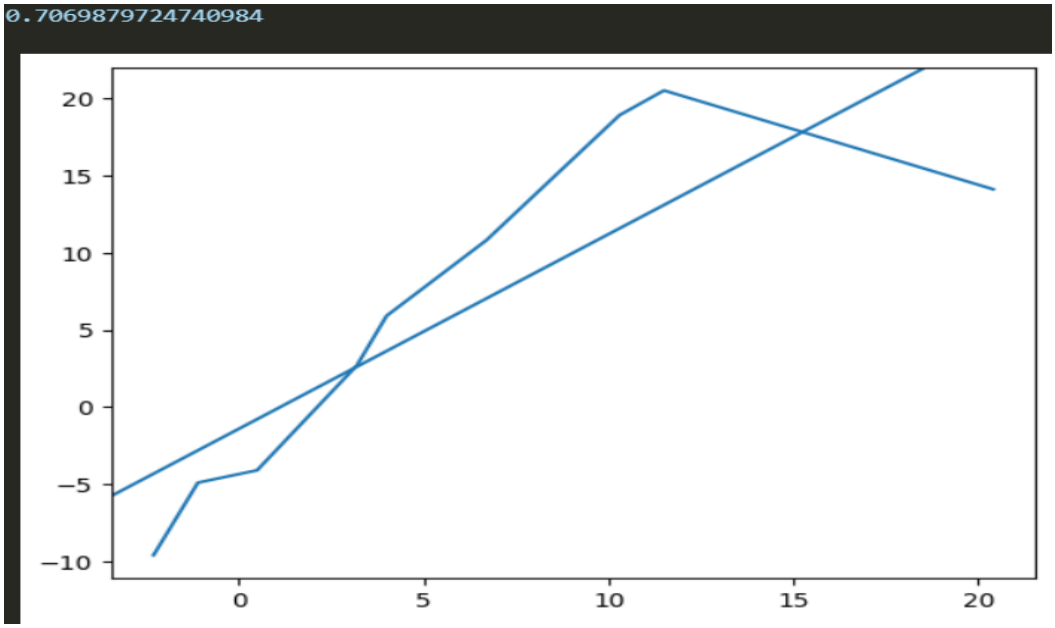
3- در این بخش در هر 4 نمونه نمودار واقعی داده ها را رسم کردیم و سپس نمودار خطی رگرسیون هر یک را نیز به همراه ضریب تعیین هرکدام نمایش میدهیم. نمودار خط راست مربوط به رگرسیون و نمودار دارای شکستگی مربوط به نمودار خود داده میباشد و ضریب تعیین هر رگرسیون نیز در بالای نمودار مربوطه قابل مشاهده است:



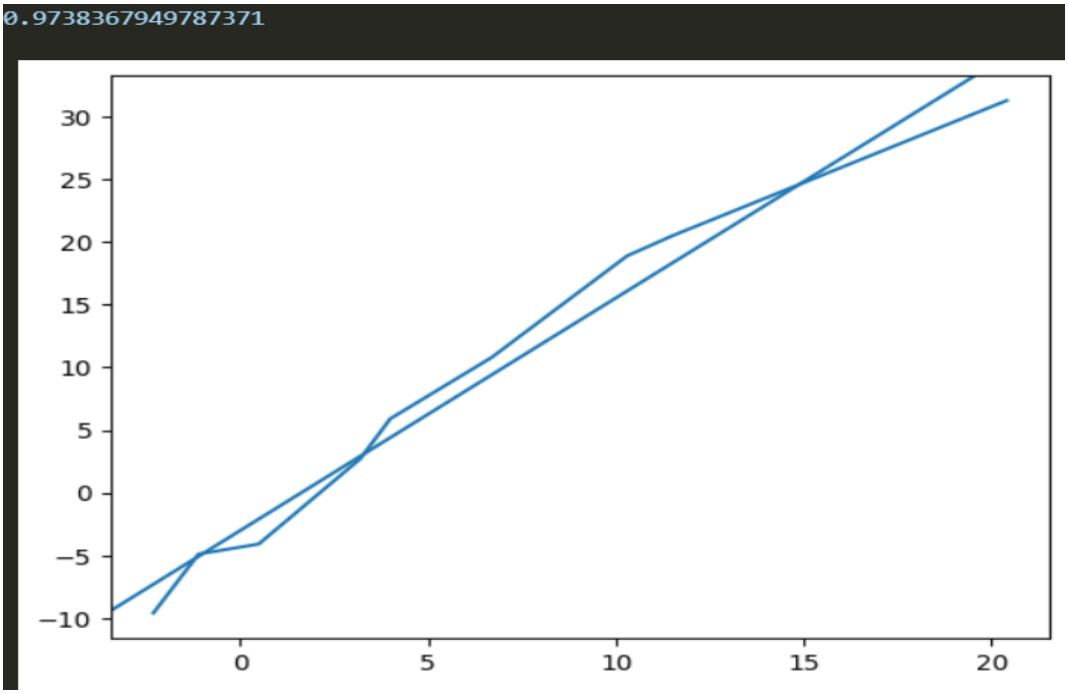
رگرسیون بر پایه داده اصلی.



رگرسیون بر پایه داده اصلی به همراه نقطه دور افتاده.



رگرسیون بر پایه داده اصلی به همراه نقطه اهرمی.



رگرسیون بر پایه داده اصلی به همراه نقاط دور افتاده و اهرمی.

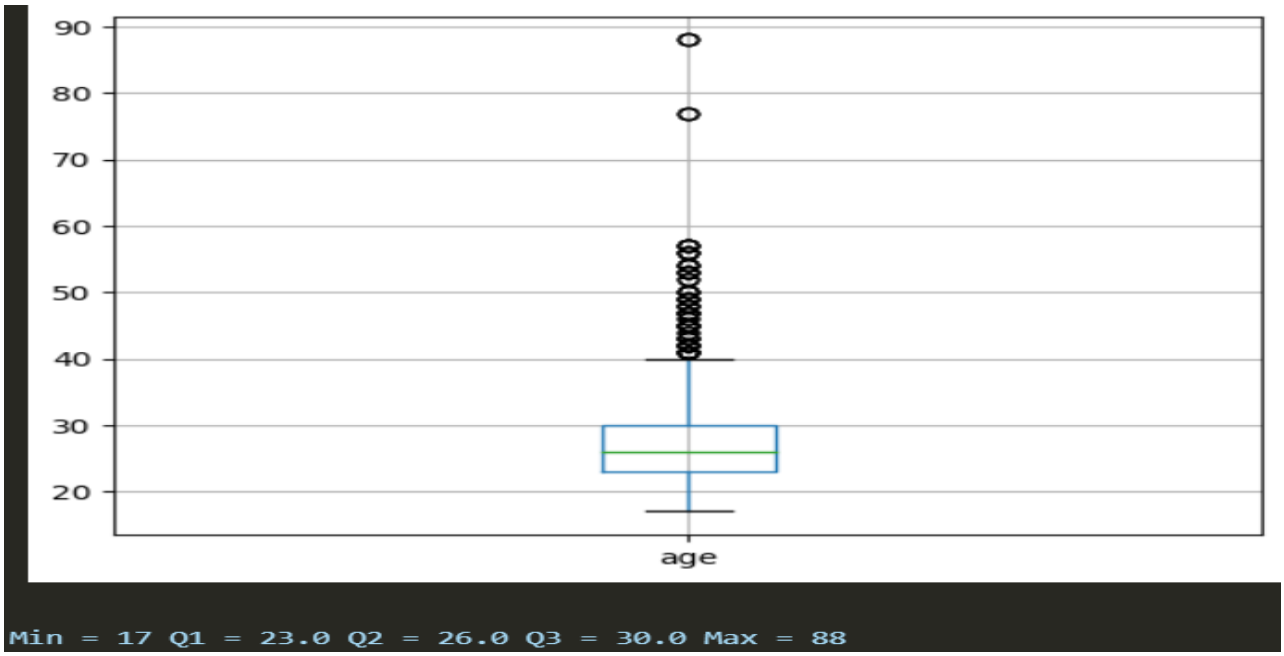
همانطور که مشخص است در نمودار اول که فاقد نقاط غیر عادی است رگرسیون بسیار نزدیک به نمودار اصلی داده ها شده است و لذا ضریب تعیین آن به 1 خیلی نزدیکتر است. همچنین تاثیر نقاط غیر عادی و اهرمی در نمودار های دوم و سوم بسیار مشهود است و باعث دور شدن رگرسیون از نمودار واقعی داده شده اند.

4- اگر حذف مستقیم این نقاط غیر عادی مقدور نباشد بهترین راهکار این است که از مدل هایی استفاده کنیم که حساسیت کمتری به این نقاط غیر عادی دارند. مثلا استفاده از مدل برآورد M یا همان روش درست نمایی بیشینه ML که به این نوع نقاط حساسیت بالایی ندارند و صرفا با عملیات های ریاضی به تخمین نسبتا خوبی میرسند.

سوال 3)

1- بهترین راهکار آن است که ابتدا تمامی این مقادیر NA و icons را یافته و سپس آن ها را با میانگین هر ستون عوض کنیم. برای همین ابتدا مقادیر NA را به NaN تبدیل کردیم تا راهکار بالا عملی باشد. در نتیجه تمامی مقادیر غیرقابل قبول در هر ستون برابر با میانگین همان ستون شدند.

2- در این بخش نمودار جعبه ای داده های مربوط به ستون age را رسم کردیم و مقادیر Min و Q1 و Q2 و Q3 و Max به صورت زیر شدند:

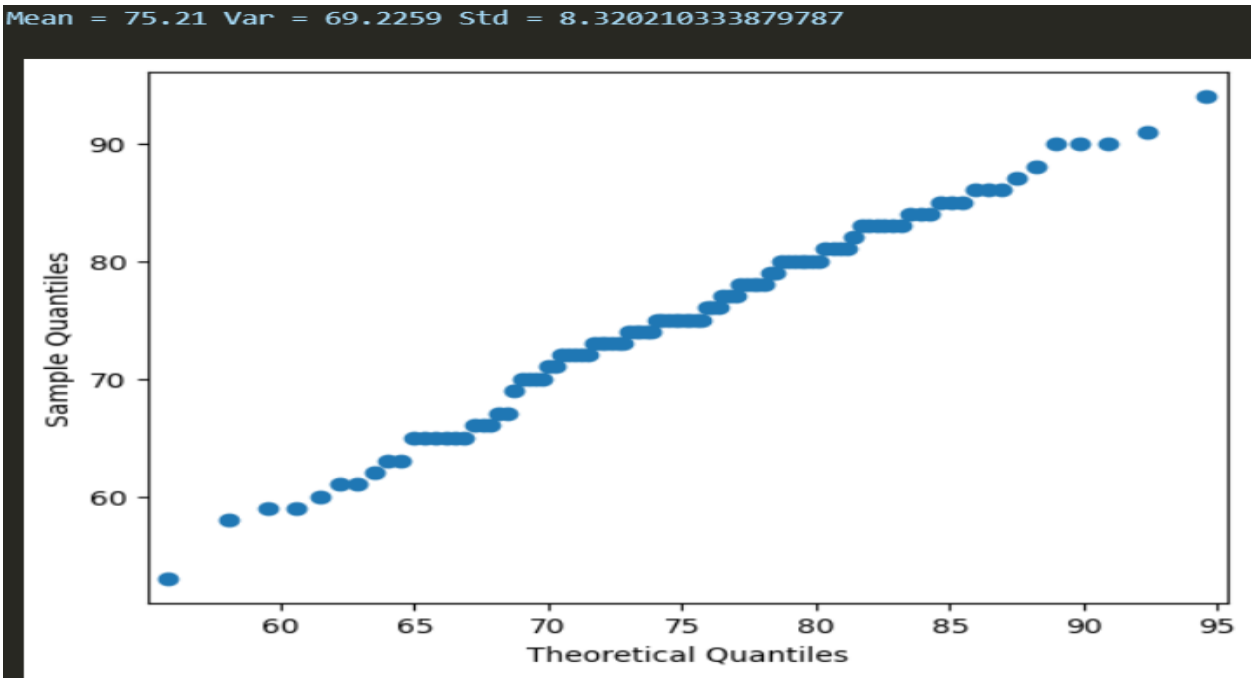


Min = 17 یعنی کمترین سن برابر 17 و Max = 88 یعنی بیشترین سن برابر 88 میباشد. Q1 = 23 و Q2 = 26 و Q3 = 30 به ترتیب یعنی چارک اول و چارک دوم (میانه) و چارک سوم برابر 23 و 26 و 30 میباشد. یعنی حدود نصف افراد بالای 26 سال و نصف افراد پایین 26 سال سن دارند. همچنین با توجه به شکل نیز میتوان گفت سن افراد تراکم بیشتری در بازه 26 تا 30 دارد. داده 88 یعنی Max نیز یک داده پرت به شمار می آید.

3-

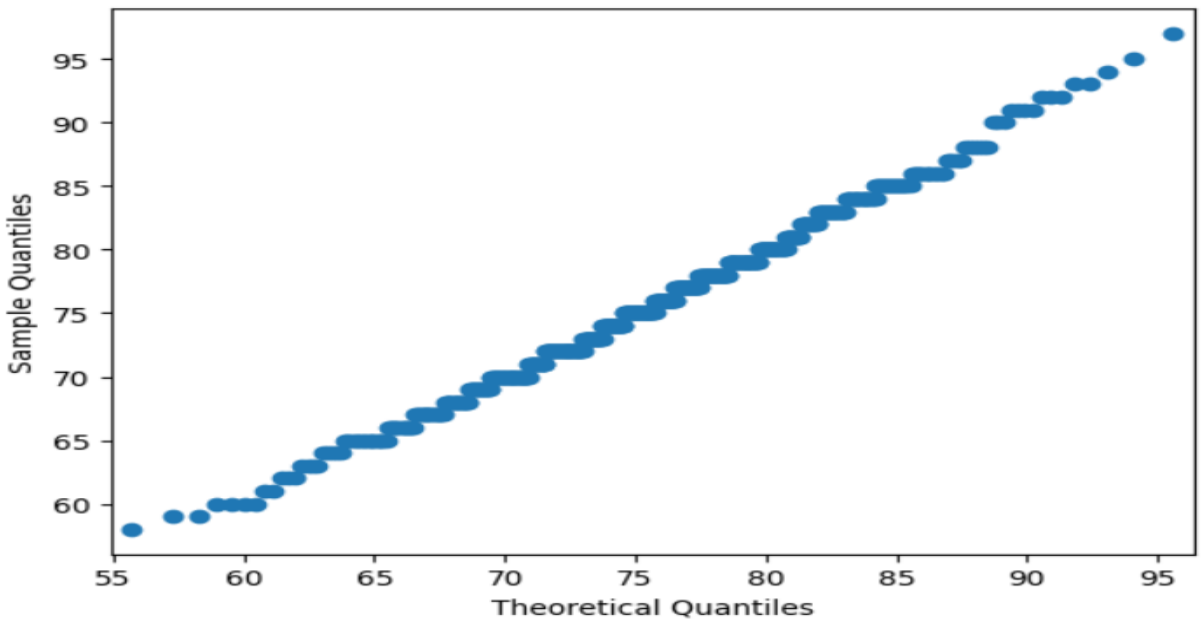
ب) نمودار Q-Q برای مقایسه توزیع دو داده به کار میرود که بفهمیم آیا دو توزیع تقریبا از یک جنس هستند یا نه. اگر توزیع این دو داده یکسان باشد مقادیر این نمودار روی نیمساز y = x قرار میگیرند و هرچه رابطه خطی در این نمودار بیشتر باشد و با نیمساز انطباق بیشتری داشته باشد یعنی دو توزیع بیشتر به یکدیگر نزدیک هستند.

الف و ج و ه) در این قسمت برای هر 3 مقدار گفته شده برای n ابتدا میانگین و واریانس و انحراف معیار نمونه های تصادفی را پیدا کردیم و نمودار Q-Q آن ها را رسم کردیم:



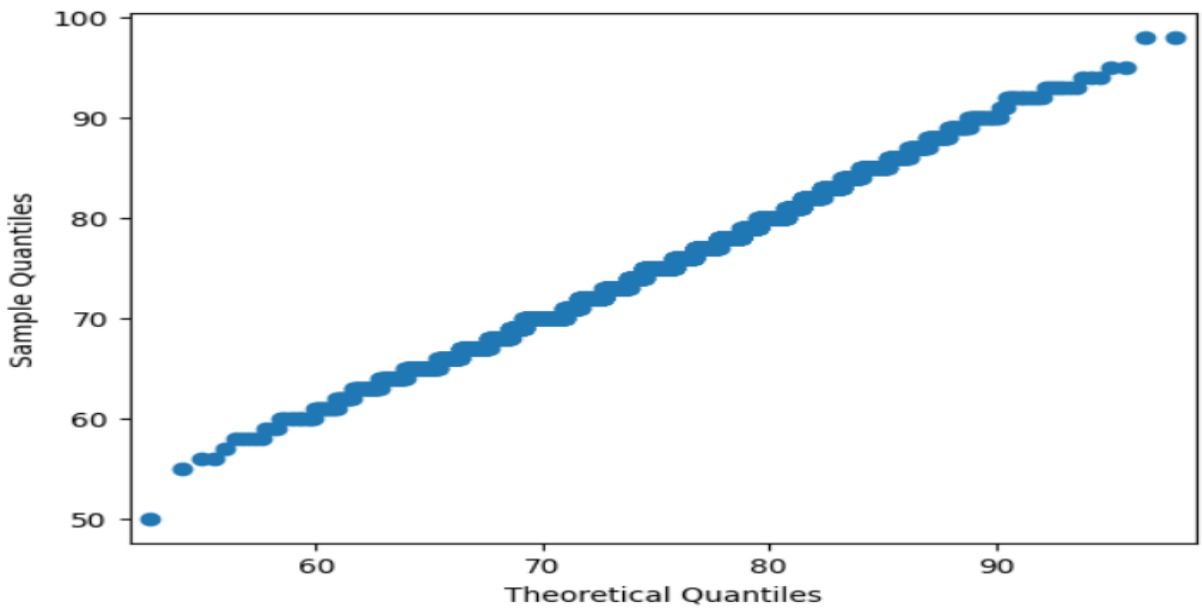
برای n = 100 نقاط نمودار تا حدودی منطبق بر یک خط راست هستند و روی نیمساز قرار میگیرند. لذا تقریبا میتوان گفت که توزیع نزدیک به نرمال است.

```
Mean = 75.66 Var = 47.976400000000005 Std = 6.926499837580306
```



برای n = 500 رابطه خطی بودن و روی نیمساز بودن این نقاط بیشتر مشهود است و تا حدودی به توزیع نرمال نزدیک میباشد.

```
Mean = 75.292 Var = 47.231736000000005 Std = 6.872534903512677
```



برای n = 2000 نیز این رابطه خطی بودن به اوج خود میرسد و میتوان گفت توزیع نرمال است.

در انتها میتوان نتیجه گرفت که هرچقدر اندازه سایز نمونه یعنی مقدار n را افرایش دهیم میزان خطی بودن و انطباق روی نیمساز در نمودار Q-Q بیشتر میشود و لذا توزیع نمونه ما به توزیع نرمال نزدیکتر خواهد شد. در نتیجه در اینجا به درستی قضیه حد مرکزی میتوان پی برد که با میل دادن n به سمت بی نهایت توزیع ما نرمال خواهد بود.

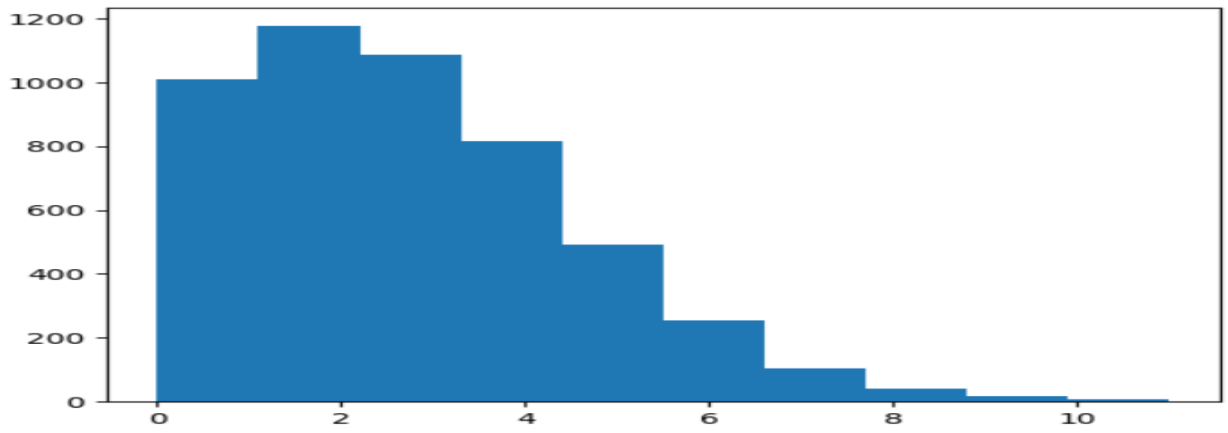
د) برای نمونه ای با اندازه n = 100 که ابتدا بررسی کردیم از آزمون شاپیرو نیز استفاده میکنیم تا نرمال بودن یا نبودن را تست کنیم. با بدست آوردن مقدار p-value فرضیه خود را آزمایش کردیم:

```
p_value = 0.5747080445289612
Normal
```

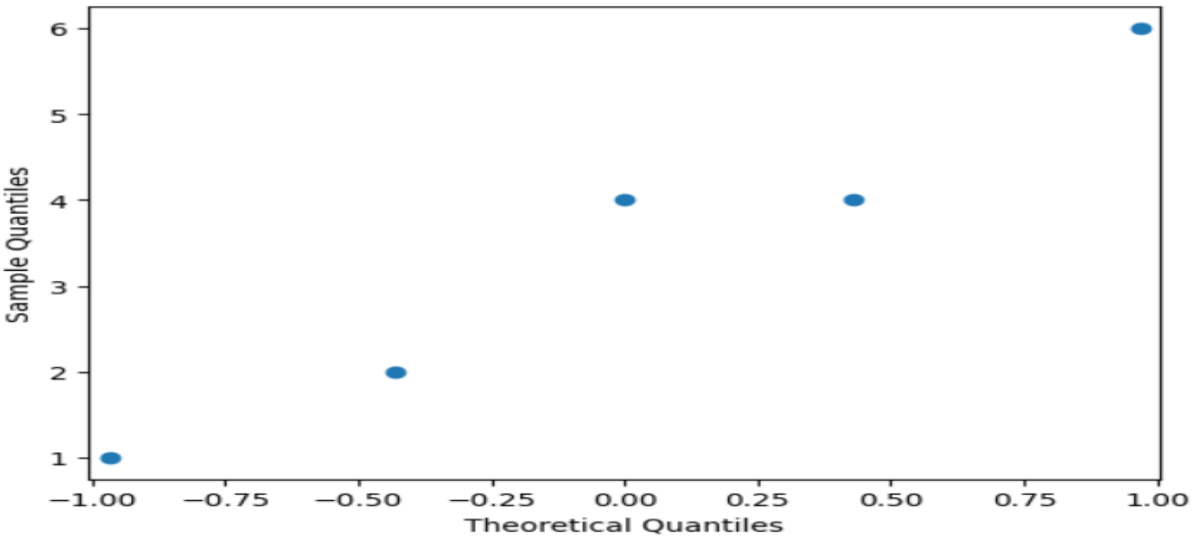
همانطور که معلوم است مقدار p-value از 5% بیشتر بوده و میتوان نتیجه گرفت که توزیع نرمال است.

4-

الف) هیستوگرام یک نمونه تصادفی به اندازه n = 5000 با توزیع پوآسون و پارامتر 3 را رسم کردیم:

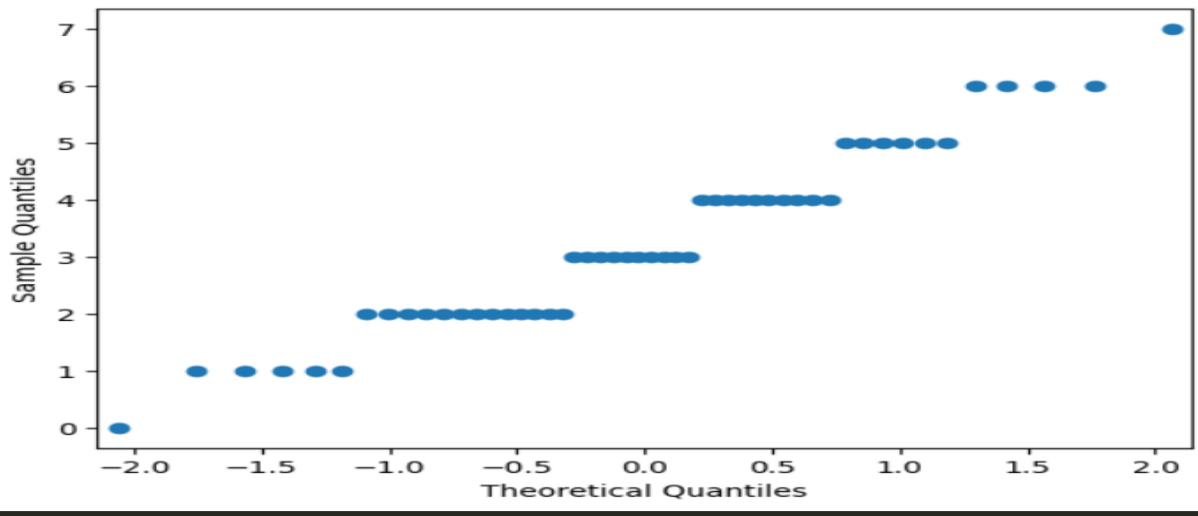


ب) در این بخش برای هر سه اندازه n که گفته شده نمودار Q-Q و همینطور p-value مربوطه را بدست آوردیم و نرمال بودن یا نبودن را بررسی کردیم:



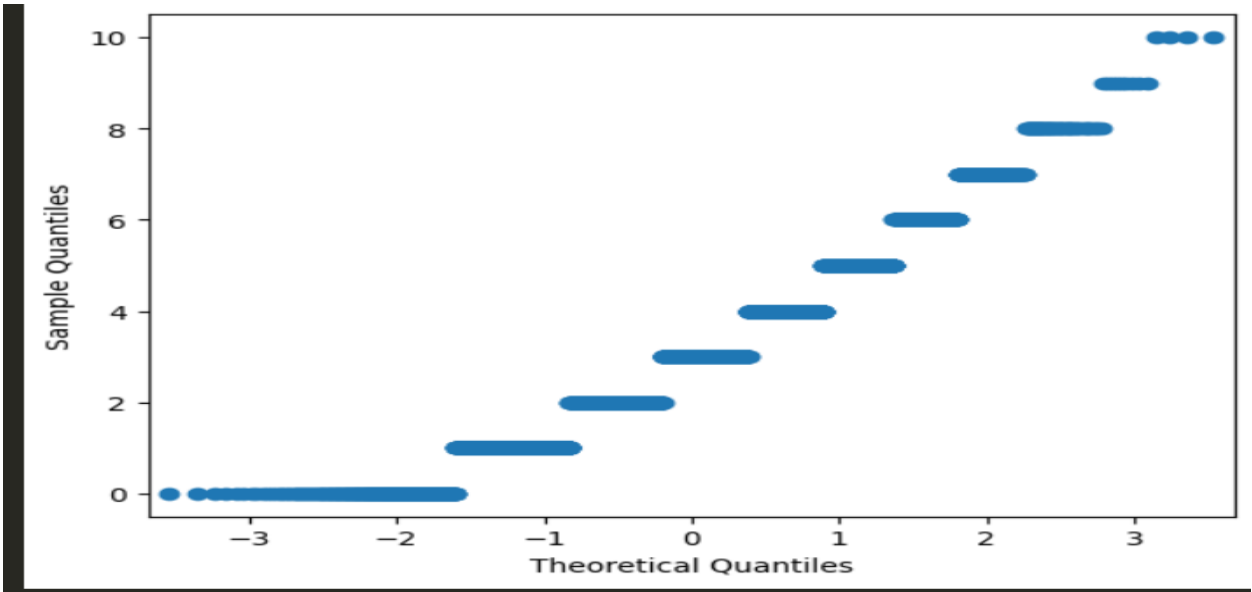

```
p_value = 0.7583119869232178
Normal
```

برای n = 5 توزیع تا حدودی به نرمال نزدیک است و پوآسون نیست. همانطور که مقدار بزرگ p-value به ما میگوید.




```
p_value = 0.041139960289001465
Not Normal
```

برای n = 50 توزیع از حالت خطی فاصله گرفته و دیگر نرمال نیست و تا حدودی به خود توزیع پوآسون نزدیک است. همچنین مقدار p-value کاهش یافته است.




```
p_value = 8.050083569038249e-38
Not Normal
```

برای n = 5000 نیز توزیع کاملا از حالت خطی و نرمال خارج شده و کاملا پوآسون است. مقدار p-value هم به شدت کاهش یافته و از نرمال نبودن توزیع خبر میدهد.

در نتیجه همانطور که مشاهده کردیم با افزایش اندازه نمونه یعنی n توزیع ما از حالت نرمال خود خارج شده و به توزیع اصلی خود یعنی پوآسون نزدیکتر شده است. همچنین مقدار p-value نیز مطابق انتظار کاهش یافته است. این مشاهدات برخلاف قضیه حد مرکزی میباشند که باید با افزایش n توزیع به نرمال نزدیکتر میشد. اما طبق انتظار چون داده های ما پوآسون بودند با افزایش مقدار n قطعا باید مقدار p-value کاهش یابد و از توزیع نرمال فاصله بگیریم و به توزیع اصلی خود یعنی پوآسون برسیم که درست است.