سوال 1:

**1.**

تشخیص ناهنجاری:

در آمار، نقاط پرت یا ناهنجاری‌ها، نقاط داده‌ای هستند که به جمعیت مشخصی تعلق ندارند. این یک مشاهده غیرطبیعی است که با مقادیر دیگر فاصله دارد. یک ناهنجاری، مشاهده‌ای است که از داده‌های خوب ساختار یافته جدا می‌شود.

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 هوشمندانه تری نیاز خواهید داشت.

حذف نویز از داده‌ها (نظیر تصاویر یا صدا):

دینویزینگ (Denoising)، فرایند حذف نویز از یک سیگنال است. این سیگنال می‌تواند یک تصویر، صدا یا یک سند باشد. می‌توان یک شبکه خودرمزگذار را به منظور یادگیری نحوه حذف نویز از تصاویر، آموزش داد.

رنگ آمیزی تصاویر

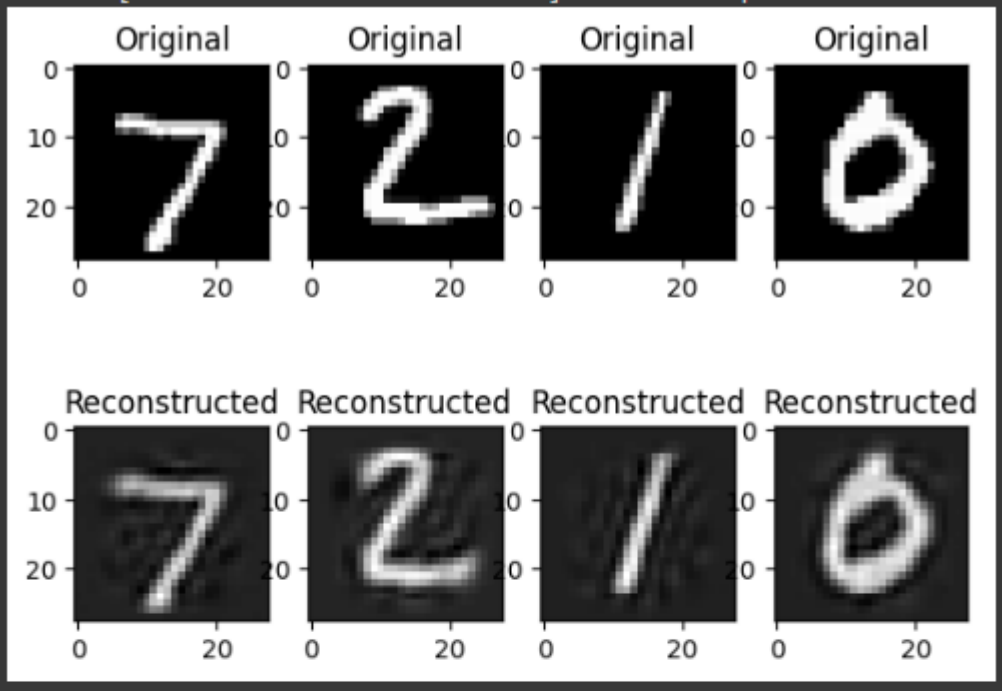
بازیابی اطلاعات

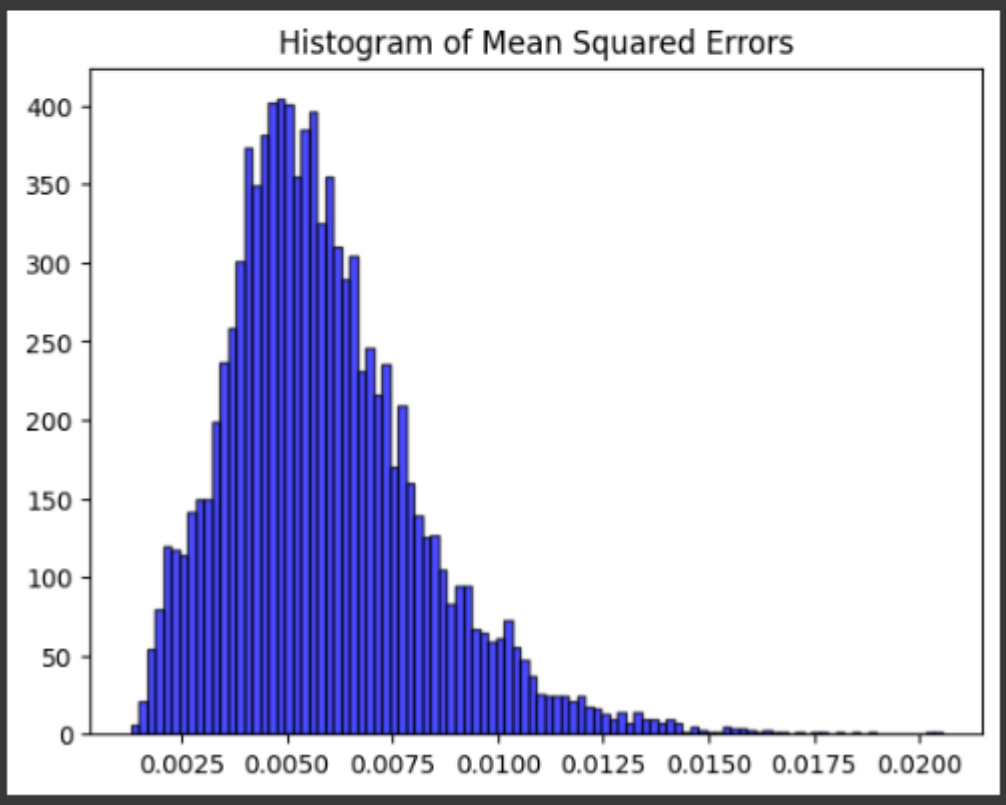
**2.** فضای پنهان نمایش فشرده داده های ورودی در اتوانکودر است. اگر فضای پنهان خیلی کوچک باشد یا به اندازه کافی بزرگ نباشد، ممکن است تمام جزئیات مهم داده های ورودی را ثبت نکند و در نتیجه بازسازی های تار را ایجاد کند. اتوانکودر تلاش می کند تا تنوع و پیچیدگی تصاویر اصلی را در یک فضای پنهان محدود نشان دهد. برای رفع تاری در بازسازی ها، می توان افزایش اندازه فضای پنهان یا استفاده از معماری های پیشرفته تر را در نظر گرفت.

(برای بخش آخر این سوال خواسته شده که براساس p_value بگوییم که داده های mse از توزیع نرمال پیروی می کند یا خیر. که با توجه به نتایج بدست آمده میتوان گفت از توزیع نرمال پیروی نمی کند(در کد مشخص است))

خروجی سوال اول. (در vs code ران نمیشد در google colab خروجی گرفتم)

```
313/313 [==============================] - 2s 5ms/step
```




```
The MSE data does not follow a normal distribution.
```

سوال 2:

**1.** نقاط پرت، نقاط داده ای هستند که با سایر نقاط داده فاصله دارند. به عبارت دیگر، آنها مقادیر غیرعادی در یک مجموعه داده هستند. نقاط پرت برای بسیاری از تحلیل‌های آماری مشکل‌ساز هستند، زیرا می‌توانند باعث شوند که آزمایش‌ها یافته‌های مهم را از دست بدهند یا نتایج واقعی را تحریف کنند. این نقاط می توانند به طور نامتناسبی بر معادله رگرسیون تأثیر بگذارند و آن را در جهت خود بکشند.

نقاط اهرمی بالا در رگرسیون خطی، نقاطی هستند که دارای مقادیر متغیر مستقل بسیار غیرمعمول در هر جهت از میانگین (بزرگ یا کوچک) هستند. چنین نکاتی قابل توجه هستند زیرا پتانسیل اعمال "کشش" یا اهرم قابل توجهی را بر روی بهترین خط مدل دارند.

(اهرم اندازه‌گیری است که نشان می‌دهد مقادیر متغیر مستقل یک مشاهده چقدر از مشاهدات دیگر فاصله دارند. نقاط اهرمی بالا، در صورت وجود، با توجه به متغیرهای مستقل، پرت هستند.)

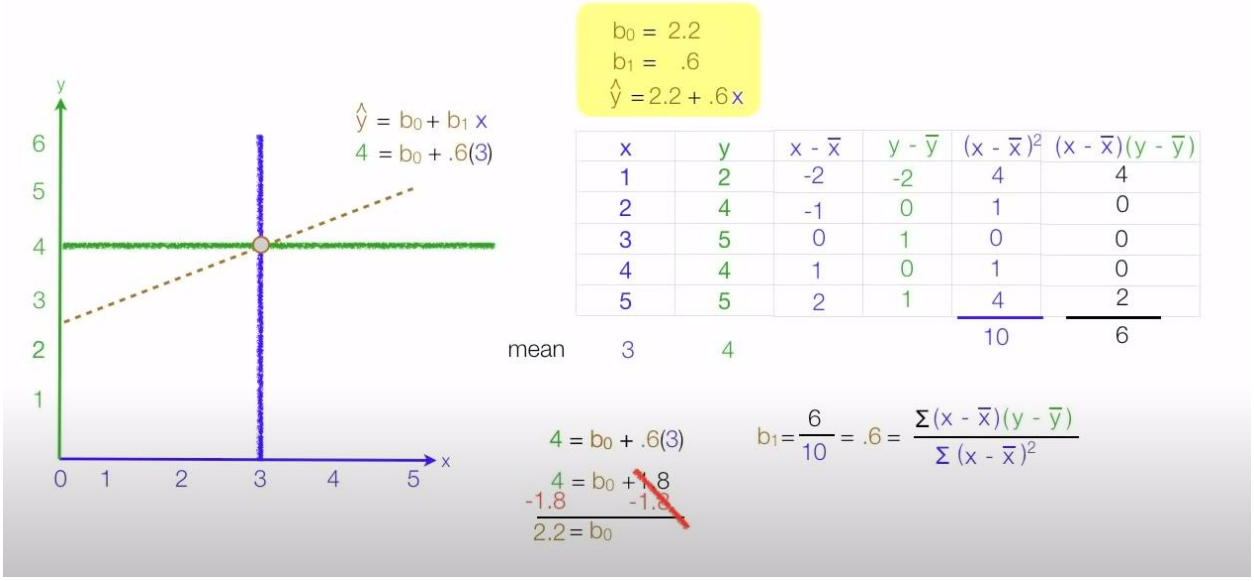
نقاط تأثیرگذار ،مشاهداتی هستند که تأثیر بسزایی در معادله رگرسیون دارند. این نقاط می توانند با تأثیر بر ضرایب برآورد شده و در نتیجه پیش بینی های مدل تأثیر بگذارند.

وجود نقاط پرت، نقاط اهرمی یا نقاط تاثیرگذار می تواند اثرات منفی بر مدل رگرسیون داشته باشد. آنها ممکن است منجر به تخمین پارامترهای مغرضانه، افزایش تنوع در پیش‌بینی‌ها و کاهش تفسیرپذیری مدل شوند. برای کاهش تأثیر این نقاط، بررسی کامل داده‌ها، شناسایی و درک ویژگی های این نقاط، و در نظر گرفتن راههای مناسب میتواند مورد توجه قرار گیرد.

**2.**ضریب تعیین یا $R^2$ میزان ارتباط خطی بین دو متغیر را اندازه گیری می کند. $R^2$ نسبت تغییرات متغیر وابسته را که می توان به متغیر مستقل نسبت داد اندازه گیری می کند. در تعاریف موجود به $R^2$، ضریب تعیین یا ضریب تشخیص نیز گفته می شود. به بیان ساده می توان گفت ضریب تعیین نشان می دهد که چند درصد تغییرات متغیر های وابسته در یک مدل رگرسیونی با متغیر مستقل تبیین می شود. به عبارت دیگر، ضریب تشخیص نشان می دهد که چه میزان یا مقدار از تغییرات متغیر وابسته مساله تحت تاثیر متغیر مستقل مساله بوده است. همچنین تا چه حدی مابقی تغییرات متغیر وابسته مساله مربوط به سایر عوامل موجود در مساله است.

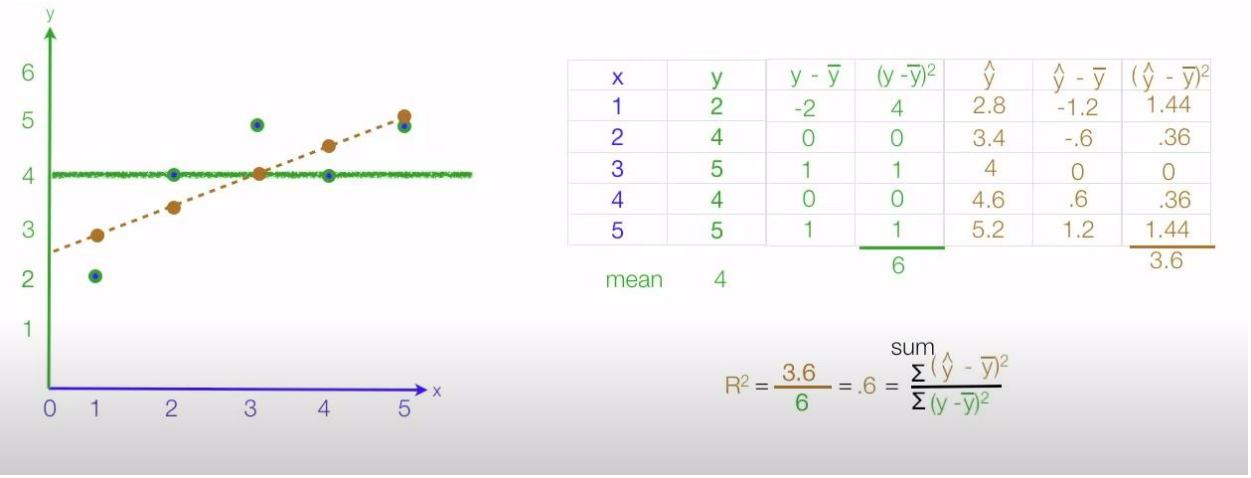
همبستگی قدرت رابطه بین یک متغیر مستقل و وابسته را توضیح می دهد، ضریب تعیین یا ضریب تشخیص بیانگر این است که تا چه اندازه واریانس یک متغیر واریانس متغیر دوم را توضیح می دهد. ضریب تعیین نمی تواند تعیین نماید که آیا مدل برازش شده دارای شیب است یا نه و به همین دلیل باید نمودارهای باقیمانده را مورد ارزیابی قرارداد.

**3.** روش پیدا کردن معادله خط




| x | y | x - x̄ | y - ȳ | (x - x̄)² | (x - x̄)(y - ȳ) |
|---|---|---|---|---|---|
| 1 | 2 | -2 | -2 | 4 | 4 |
| 2 | 4 | -1 | 0 | 1 | 0 |
| 3 | 5 | 0 | 1 | 0 | 0 |
| 4 | 4 | 1 | 0 | 1 | 0 |
| 5 | 5 | 2 | 1 | 4 | 2 |
| | | | | 10 | 6 |

روش پیدا کردن ضریب تعیین یا $R^2$




| x | y | $y - \bar{y}$ | $(y - \bar{y})^2$ | $\hat{y}$ | $\hat{y} - \bar{y}$ | $(\hat{y} - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 2 | -2 | 4 | 2.8 | -1.2 | 1.44 |
| 2 | 4 | 0 | 0 | 3.4 | -.6 | .36 |
| 3 | 5 | 1 | 1 | 4 | 0 | 0 |
| 4 | 4 | 0 | 0 | 4.6 | .6 | .36 |
| 5 | 5 | 1 | 1 | 5.2 | 1.2 | 1.44 |
| mean | 4 | | 6 | | | 3.6 |

$$R^2 = \frac{3.6}{6} = .6 = \frac{\text{sum}\ \Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

**4.** یک راه ساده میتواند در نظر نگرفتن این نقاط در تحلیل های خود باشد. برای افزایش مقاومت مدل رگرسیون خطی در برابر نقاط پرت، استفاده از تکنیک های رگرسیون قوی مانند رگرسیون Huber یا RANSAC را در نظر گرفت که کمتر تحت تأثیر نقاط داده شدید قرار می گیرند. علاوه بر این، استفاده از مقیاس‌بندی ویژگی و نرمال‌سازی می‌تواند تأثیر نقاط پرت را کاهش دهد.

سوال 3:

**1.** برای missing value handling چند روش وجود دارد. اول data understanding. یعنی باید چک کینم که nan بودن اون داده منظور دار است یا نه. برای مثال یک دیتاست راجب کارکنان یک کارخانه هستند و این کارخانه شعبه های مختلفی داره و یک ستون از داده ها مربوط به آدرس شعبه است. منتهی یک سری کارکنان این کارخانه آنلاین کار میکنند و شعبه کار کردن ندارند. برای همین nan زده شده. در این موقع باید این دیتا ها را صرفا پیدا کنیم و مثلا دستی بذاریم 0 یا یک کلمه خاصی بذاریم چون معنا دار nan شده اند.

مثلا در همین دیتاست fifa(برای دروازه بانها pace، nanهست) ممکن است pace خیلی برای دروازه بان تعریف شده نباشد برای همین nan زده شده است(برای همه دروازه بانها) پس میتوان با گذاشتن 0 یا 1 یا ... برای همشون بشه مشکل را حل کرد.

یک سری دیتا هایی هستند که میشه با مقدار ثابت پر کرد البته این مقادیر با تحلیل داده ها باشد. باید بدونیم که این مقادیر باید معنایی برای اون داده داشتته باشد. روشهای دیگری مانند روشهای آماری وجود دارد که میتوان این missing value ها را با مقدار میانگین یا مد یا میانه پر کنیم که این از روشهای ساده ولی کارساز به شمار میرود.ولی خب باید معنا دار باشد. یا روشهای دیگری مانند اینکه بیایم این دادها را با توجه داده های بالاتر یا پایین تر از خودشان در دیتاست پر کنیم. بیشتر برای داده های مربوط به زمان بکار میرود. یا اصلا میتوان کل اون داده را حذف کرد😊 (در اینجا معیار این نیست) یا روشهای خیلی خیلی زیاد دیگر.

در این کیس خودمون میتوان مقدار میانگین یا میانه را برای داده های miss شده در ستونهای pace, dribbling استفاده کرد.(من از میانگین استفاده کردم)

**2.** در اینجا منظور از مقدار min، جوان ترین بازیکن است و منظور از مقدار max، مسن ترین بازیکن.

منظور از Q1 = سنی که یک چهارم بازیکنان از آن جوانتر هستند

منظور از Q2 = سنی که نیمی از بازیکنان از آن جوانتر هستند

منظور از Q3 = سنی که یک چهارم بازیکنان از آن بزرگتر هستند

**3.** Q-Q plot یک ابزار گرافیکی است که در آمار برای ارزیابی اینکه آیا یک مجموعه داده از توزیع نظری خاصی پیروی می کند یا خیر استفاده می شود. این نمودار چندک های داده های مشاهده شده را با چندک های توزیع نظری انتخاب شده مقایسه می کند.

برای بخش ج: با توجه به نتایج بدست آمده(p_value , Q-Q plot) به این نتیجه میتوان رسید که سن بازیکنان از یک توزیع نرمال پیروی میکند (در کد مشخص است)

برای بخش ه: می‌توانیم ببینیم که با افزایش تعداد نمونه‌ها، p_value بسیار کوچک‌تر از 0.05 می‌شود و اگرچه با n کوچک‌تر به دیدیم که احتمالاً فرضیه درست است و وزن بازیکنان از توزیع نرمال پیروی میکند، اما برای n های بزرگ می‌توانیم این فرضیه را رد کنیم. (در اینجا هم طبق CLT باید برای n های بزرگ به توزیع نرمال میل میکرد که اینگونه نشده که جلوتر به یک چیزهایی اشاره میکنم)

**4.** از نمودار Q-Q می‌توان نتیجه گرفت که داده‌ها تقریبا به طور نرمال توزیع می‌شوند، اگرچه از p_value محاسبه‌شده از آزمون Shapiro-Wilk فرضیه صفر رد می‌شود و میگوید که داده‌های آزمایش‌شده به طور نرمال توزیع نشده‌اند(برای n=5 میگوید توزیع نرمال میشود درصورتی که خیلی هم نیست ولی برای n = 50, 500 میگوید توزیع نرمال نمیشود).

از CLT می دانیم که هر نمونه ای از داده ها به اندازه کافی بزرگ تمایل به پیروی از توزیع نرمال دارد، بنابراین در اینجا میتوان گفت که آزمون Shapiro-Wilk نمیتواند درست پیشبینی کند.

بعد از سرچ کردن: اگر حجم نمونه شما بزرگ باشد، حتی انحرافات خفیف از نرمال بودن توسط آن تست ها رد می شود. 😐

تست ها اشتباه میکنند پس.