

# گزارش اول درس کلان داده ها

موضوع

بررسی اضطراب برای دانشجویان در هنگام امتحان به کمک نشانگرهای زیستی

پارسا اسدnezhad

Parsaasadnezhad2@gmail.com

## فهرست

4 .....	مقدمه
4 .....	به زبان ساده
4 .....	نتایج کلیدی
5 .....	نقش و دستاوردهای کلیدی این مطالعه:
5 .....	چالش‌های مهم شناسایی شده:
5 .....	چرا این مطالعه مهم است؟
6 .....	مسیر پیش رو برای تحقیقات آینده
6 .....	بیومارکرهای دیجیتال (Digital Biomarkers)
6 .....	مثال‌هایی از بیومارکرهای دیجیتال
7 .....	چالش‌ها
7 .....	نکات مهم
7 .....	جمع‌بندی
7 .....	دیتاست
7 .....	ساختار داده‌ها
8 .....	فایل‌های موجود در هر پوشه امتحان
8 .....	نکات مهم زمانی
8 .....	مشخصات فنی
8 .....	اطلاعات شرکت‌کنندگان
8 .....	ساختار داده‌ها
11 .....	تحلیل:
11 .....	تعداد سلول‌های دیتاست
16 .....	اطلاعات کلی
17 .....	ستون‌ها
17 .....	توضیحات آماری

18 .....	تحليل
18 .....	گیتهاب

9 .....	شکل 1 پوشه مربوط به تمام دانشجویان .....
9 .....	شکل 2 پوشه مربوط به هر دانشجو .....
9 .....	شکل 3 دیتاستها مربوط به هر امتحان .....
10 .....	شکل 5 تعداد سلول برخی از دیتاستها.....
10 .....	شکل 6 یک دیتا فریم شامل تمامی دیتاستها.....
12 .....	شکل 7 تعداد سلول‌ها به تفکیک دانشجویان .....
13 .....	شکل 8 اطلاعات تمامی دیتاستها برای دانشجو s1 .....
14 .....	شکل 9 نمودار میله‌ای برای تعداد سلول‌ها در تمامی دیتاستها برای دانشجو 10 .....
14 .....	شکل 10 نمودار pie برای تعداد سلول‌ها در تمامی دیتاستها برای دانشجو 10 .....
15 .....	شکل 11 نمودار میله‌ای برای تعداد سلول‌ها در تمامی دیتاستها برای دانشجو 7 .....
16 .....	شکل 12 نمودار pie برای تعداد سلول‌ها در تمامی دیتاستها برای دانشجو 7 .....
16 .....	شکل 13 اطلاعات کلی .....

## مقدمه

این تحقیق بررسی می‌کند که آیا با استفاده از دستگاه‌های پوشیدنی مثل ساعت‌های هوشمند می‌توان عملکرد دانشجویان در امتحان را پیش‌بینی کرد یا نه. پژوهشگران در طول امتحانات، اطلاعاتی مثل میزان تعریق و دمای بدن دانشجویان را اندازه‌گیری کردند.

آنها متوجه شدند دانشجویانی که الگوهای خاصی در تعریق (استرس پیشتر) داشتند، معمولاً نمرات پایین‌تری می‌گرفتند. البته این تحقیق هنوز در مراحل اولیه است و عوامل دیگری مثل حرکت‌های دانشجویان حین امتحان ممکن است روی دقت داده‌ها تأثیر بگذارد.

هدف این مطالعه این بود که با استفاده از دستگاه‌های پوشیدنی، داده‌های فیزیولوژیکی (مثل هدایت پوستی و دمای پوست) از ۰۱ دانشجو در طول سه امتحان جمع‌آوری شود. سپس، پژوهشگران از داده‌های هدایت پوستی (که نشان‌دهنده استرس است) برای پیش‌بینی نمرات دانشجویان (بالا یا پایین) استفاده کردند.

## به زبان ساده

- دانشجوها در حین امتحان، دستگاه‌هایی مثل ساعت هوشمند یا سنسور پوستی پوشیدند.
- این دستگاه‌ها میزان تعریق (که با هدایت پوستی اندازه‌گیری می‌شود) و دمای پوست را ثبت کردند.
- بعد از امتحان، دانشمندان بررسی کردند که آیا بین الگوی استرس (تفییرات تعریق) و نمره امتحان ارتباطی وجود دارد یا نه.

پژوهشگران ابتدا سیگنال‌های هدایت پوستی (که نشان‌دهنده استرس است) را فیلتر و پردازش کردند تا الگوهای کلی تغییرات استرس در طول امتحان به دست آید. سپس با استفاده از این الگوها، یک مدل هوش مصنوعی آموزش دادند تا پیش‌بینی کند که آیا نمره دانشجو بالا خواهد بود یا پایین.

## نتایج کلیدی

- دقت پیش‌بینی مدل: بین ۷۰ تا ۸۰٪ — یعنی در بیشتر موارد توانسته نمره را به درستی حدس بزند.
- کاربرد عملی: این نشان می‌دهد که دستگاه‌های پوشیدنی (مثل ساعت‌های هوشمند) پتانسیل پیش‌بینی عملکرد افراد در شرایط پر استرس (مثل امتحان) را دارند.
- نمودار میانگین استرس: محققان همچنین میانگین تغییرات سطح استرس دانشجویان را در طول امتحان ترسیم کردند که به‌طور کلی روند افزایش و کاهش استرس را نشان می‌دهد.

## نقش و دستاوردهای کلیدی این مطالعه:

این تحقیق دو سهم اصلی به حوزه علوم داده و سلامت دیجیتال اضافه می‌کند:

### ۱. تهیه یک مجموعه داده منحصر به فرد از استرس امتحان در شرایط واقعی

- برای اولین بار، داده‌های فیزیولوژیکی دانشجویان (مانند هدایت پوستی و دمای بدن) با استفاده از دستگاه‌های پوشیدنی در طول امتحان جمع آوری شد.
  - این داده‌ها به صورت عمومی منتشر خواهند شد تا سایر پژوهشگران بتوانند از آن برای تحقیقات مرتبط (مثل استرس، عملکرد شناختی یا یادگیری ماشین) استفاده کنند.
- ### ۲. تحلیل اولیه ارتباط بین استرس و نمره امتحان
- پژوهشگران نشان دادند که الگوهای هدایت پوستی (تعریق ناشی از استرس) تا حدی می‌تواند نمره دانشجو را پیش‌بینی کند (با دقت ۷۰-۸۰٪).
  - این یافته‌ها گام اولیه‌ای برای توسعه ابزارهای پوشیدنی هوشمند است که مثلاً به دانشجویان هشدار می‌دهند استرسشان در حال تاثیرگذاری بر عملکردشان است.

## چالش‌های مهم شناسایی شده:

محققان به اندازه کوچک نمونه‌گیری (۱۰ نفر) به عنوان یک محدودیت اشاره کرده‌اند، اما تأکید می‌کنند که داده‌های طولی جمع آوری شده از هر شرکت‌کننده در سه امتحان مختلف (دو میان‌ترم و یک پایانی)، ارزش تحلیلی خاصی به این پژوهش داده است.

- مشکل نویز حرکتی: در محیط‌های واقعی (مثل جلسه امتحان)، حرکت‌های غیررادی کاربران (مثل تکان دادن دست) باعث اختلال در داده‌های سنسورها می‌شود.
- نیاز به روش‌های قوی‌تر: برای تشخیص دقیق‌تر هیجانات و استرس، الگوریتم‌های بهتری نیاز است که تحت تأثیر عوامل محیطی قرار نگیرند.

## چرا این مطالعه مهم است؟

- پژوهشکی: در آینده ممکن است از چنین سیستم‌هایی برای پایش استرس بیماران یا سربازان استفاده شود.
- آموزش: معلمان می‌توانند با تحلیل استرس دانشآموزان، روش‌های امتحانی بهینه‌تری طراحی کنند.

- توسعه فناوری: انتشار این داده‌ها به پیشرفت هوش مصنوعی در حوزه «سلامت دیجیتال» کمک می‌کند.

## مسیر پیش‌رو برای تحقیقات آینده

۱. افزایش حجم نمونه‌گیری: انجام آزمایش‌های گستردۀ تر با گروه‌های بزرگتر از دانشجویان
۲. بهبود فیلترهای پردازش سیگنال:

  - توسعه فیلترهای تطبیقی برای کاهش نویفه‌های حرکتی (Motion Artifacts) که کیفیت داده‌های پوشیدنی را تحت تأثیر قرار می‌دهند.

۳. روش‌های پیشرفته‌تر سنجش استرس:
  - استفاده از ترکیب چندین نشانگر فیزیولوژیکی (مثل ضربان قلب + دمای پوست + EDA) برای تخمین دقیق تر سطح استرس.

## بیومارکرهای دیجیتال (Digital Biomarkers)

به زبان ساده، بیومارکرهای دیجیتال نشانگرهای قابل اندازه‌گیری و عینی هستند که از طریق دستگاه‌های دیجیتال (مثل ساعت‌های هوشمند، اپلیکیشن‌های موبایل، یا سنسورهای پوشیدنی) جمع‌آوری می‌شوند و اطلاعاتی درباره سلامت، رفتار یا وضعیت فیزیولوژیک فرد ارائه می‌دهند.

## مثال‌هایی از بیومارکرهای دیجیتال

- ضربان قلب و تغییرات آن (برای بررسی استرس یا بیماری‌های قلبی)
- الگوی خواب (با استفاده از سنسورهای حرکتی)
- میزان فعالیت بدنی (قدمشمار یا کالری‌سوزی)
- تغییرات صدا (برای تشخیص اختلالات عصبی مثل پارکینسون)
- هدایت پوستی (GSR) (برای سنجش استرس یا هیجان)

## چالش‌ها

- دقت داده‌ها: حرکات ناخواسته (مثل تکان خوردن دست) ممکن است نتایج را تحریف کنند.
- حريم خصوصی: ذخیره و استفاده از داده‌های حساس افراد نیاز به قوانین محکم دارد.
- تفاوت‌های فردی: واکنش بدن هر فرد به عوامل استرس‌زا متفاوت است.

## نکات مهم

- این تحقیق مقدماتی است و برای استفاده عملی نیاز به توسعه بیشتر دارد.
- عواملی مثل تفاوت‌های فردی یا خطاهای سنسورها می‌توانند روی دقت تأثیر بگذارند.
- در آینده، چنین سیستم‌هایی ممکن است به دانشجویان کمک کنند تا استرس خود را در موقعیت‌های حساس مدیریت کنند.

## جمع‌بندی

این مطالعه نشان می‌دهد که ردیابی استرس با فناوری پوشیدنی می‌تواند ابزار مفیدی برای پیش‌بینی عملکرد باشد، هرچند هنوز جای بهبود دارد.

## دیتابست

### ساختار داده‌ها

شامل نمرات هر دانشجو می‌باشد: **StudentGrades.txt**

حاوی پوشش‌هایی برای هر شرکت‌کننده با نام‌های S1، S2 وغیره است: **Data.zip**

پوشش‌های هر شرکت‌کننده شامل سه پوشش می‌شود:

- "Final" (امتحان پایانی)
- '1' Midterm (میان‌ترم اول)
- '2' Midterm (میان‌ترم دوم)

## فایل‌های موجود در هر پوشه امتحان

هر پوشه امتحان شامل فایل‌های CSV زیر است:

- **ACC.csv**: داده‌های شتاب‌سنج
- **BVP.csv**: داده‌های حجم خون پالسی
- **EDA.csv**: داده‌های فعالیت الکترودرمال (هدایت پوستی)
- **HR.csv**: داده‌های ضربان قلب
- **IBI.csv**: فواصل بین ضربان‌های قلب
- **tags.csv**: برچسب‌های زمانی
- **TEMP.csv**: داده‌های دمای پوست
- **info.txt**: اطلاعات دقیق درباره هر یک از این فایل‌ها

## نکات مهم زمانی

- تمام برچسب‌های زمانی یونیکس برای عدم شناسایی تغییر تاریخ داده‌اند اما تغییر زمانی نداشته‌اند.
- تغییر تاریخ به گونه‌ای انجام شده که وضعیت ساعت تابستانی (CT/CDT) را تغییر نمی‌دهد.
- تمام امتحانات ساعت 09:00 صبح (به وقت CDT یا CT بسته به تاریخ) شروع شده‌اند.
  - مدت زمان میان‌ترم‌ها: 1.5 ساعت
  - مدت زمان امتحان پایانی: 3 ساعت

## مشخصات فنی

- فرکانس نمونه‌برداری آرایه‌ها در ساختار داده‌ها ذکر شده است.

## اطلاعات شرکت‌کنندگان

- مجموعه داده شامل 2 شرکت‌کننده زن و 8 شرکت‌کننده مرد می‌باشد.
- با این حال، جنسیت شرکت‌کنندگان به منظور حفظ حریم خصوصی و عدم شناسایی ذکر نشده است.

## ساختار داده‌ها

این مجموعه داده‌ها در ۱۰ فولدر که هر کدام برای دانشجویی خاصی است طبقه‌بندی شده است.



شکل ۱ پوشه مربوط به تمام دانشجویان

و درون هر کدام از این پوشه ها، سه پوشه دیگر شامل اطلاعات زیستی برای ۳ امتحان متفاوت جمع آوری شده است.



شکل ۲ پوشه مربوط به هر دانشجو

و در قسمت آخر، داده های جمع آوری شده را در فایل هایی با پسوند .CSV می بینیم.



شکل ۳ دیتاستها مربوط به هر امتحان

که همانطور که در بالا ذکر شد، هر کدام از فایل های بالا اطلاعات متفاوتی را در خود جای داده اند. به عنوان مثال فایل HR.csv شامل اطلاعات ضربان قلب دانشجویان در طول برگزاری امتحان بوده و با فایل TEMP.csv به اطلاعات چون دمای بدن دانشجویان اشاره می کند.

یکی از چالش هایی که در نگاه اول می توان مشاهده کرد این است که با توجه به استفاده از سنسور های متفاوت نمی توان همه داده ها رو در یک Dataframe خاص، قرار داد، چرا که اندازه یا shape هر فایل با سایر فایل های دیگر متفاوت می باشد.

به عنوان مثال:

```

S4/Midterm 1/BVP.csv: (748397, 1)
S4/Midterm 1/HR.csv: (11685, 1)
S4/Midterm 1/EDA.csv: (46777, 1)
S4/Midterm 1/TEMP.csv: (46777, 1)
S4/Midterm 1/tags.csv: (2, 1)
S4/Midterm 1/IBI.csv: (210, 2)
S4/Midterm 1/ACC.csv: (374203, 3)
S9/Midterm 2/BVP.csv: (795620, 1)
S9/Midterm 2/HR.csv: (12423, 1)
S9/Midterm 2/EDA.csv: (49729, 1)
S9/Midterm 2/TEMP.csv: (49721, 1)
S9/Midterm 2/IBI.csv: (540, 2)
...
S8/Midterm 1/tags.csv: 0 cells
S8/Midterm 1/IBI.csv: 420 cells

```

شکل ۴ تعداد سلوں برخی از دیتاستها

در شکل بالا می‌توان دید که یک دیتاست مانند BVP.csv برای امتحان Midterm 1 منحصر به دانشجوی S4 شامل 748397 سطر می‌باشد در حالی که همین دیتاست برای دانشجوی S9 شامل 795620 سطر است، که این اختلاف به علت تفاوت در جمع آوری داده‌ها در هنگام آزمایش است. همینطور دیتاست‌های متفاوت مانند HR و IBI تفاوت بسیاری دارند که این اختلاف ممکن است به علت تفاوت بنیادین در ساختار و جنس داده‌ها مرتبط باشد.

در این قسمت من برای نظم و ساختارمندی مناسبتر، اطلاعات کلیدی تمامی دیتاست‌ها را در یک دیتاست قرار دادم، که در ادامه تصویری را با هم مشاهده می‌کنیم.

```

summary_df.info()

0.0s

<class 'pandas.core.frame.DataFrame'>
Index: 191 entries, 125 to 22
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   File Name    191 non-null    object 
 1   Rows          191 non-null    int64  
 2   Columns       191 non-null    int64  
 3   Total Cells  191 non-null    int64  
dtypes: int64(3), object(1)
memory usage: 7.5+ KB

```

شکل ۵ یک دیتافریم شامل تمامی دیتاستها

خروجی summary\_df.info نشان می‌دهد که این DataFrame شامل اطلاعات زیر است:

1. تعداد کل ردیف‌ها (Index):

این DataFrame شامل 191 ردیف است که هر ردیف مربوط به یک فایل CSV است.

2. ستون‌ها (Columns):

این DataFrame دارای 4 ستون است:

- .(object) CSV (نوع داده): نام فایل CSV
- .(64int) Rows: تعداد ردیف‌های موجود در هر فایل CSV
- .(64int) Columns: تعداد ستون‌های موجود در هر فایل CSV
- .(64int) Total Cells: تعداد کل سلول‌ها (ردیف × ستون) در هر فایل CSV
- .Non-Null Count: 3

تمام ستون‌ها دارای 191 مقدار غیر تهی (non-null) هستند، به این معنی که هیچ مقداری در این DataFrame گم نشده است.

3. حجم حافظه (Memory Usage):

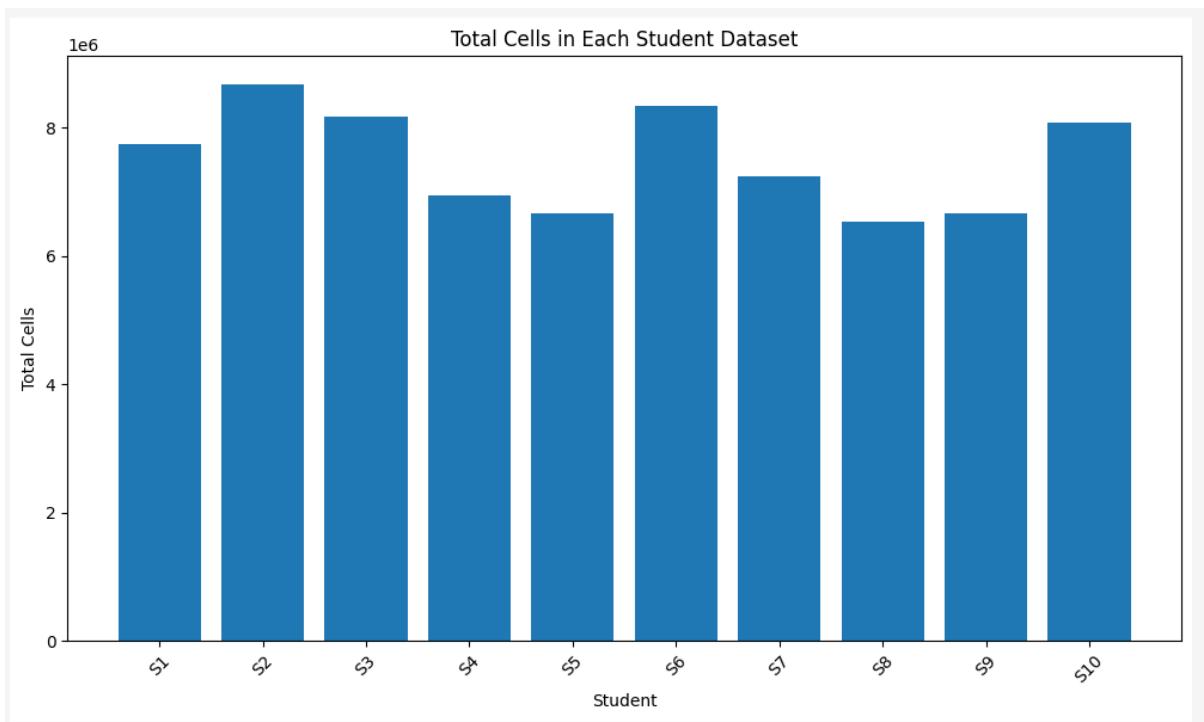
این DataFrame حدود 7.5 کیلوبایت از حافظه را اشغال می‌کند.

## تحلیل:

- این DataFrame به خوبی اطلاعات مربوط به فایل‌های CSV را خلاصه کرده است.
- هیچ داده گمشده‌ای وجود ندارد، بنابراین نیازی به پاکسازی داده‌ها نیست.
- ستون‌های عددی (Total Cells، Rows، Columns) می‌توانند برای تحلیل‌های آماری یا مصورسازی استفاده شوند.

## تعداد سلول‌های دیتابست

به کمک ساختار دیتابست بالا به راحتی می‌توان تعداد سلول‌های تمام دیتابست‌ها را محاسبه نمود. این تعداد برابر است با 75003483 که چیزی بیشتر از 75 میلیون داده است که برای این درس یعنی کلان داده‌ها مناسب خواهد بود.



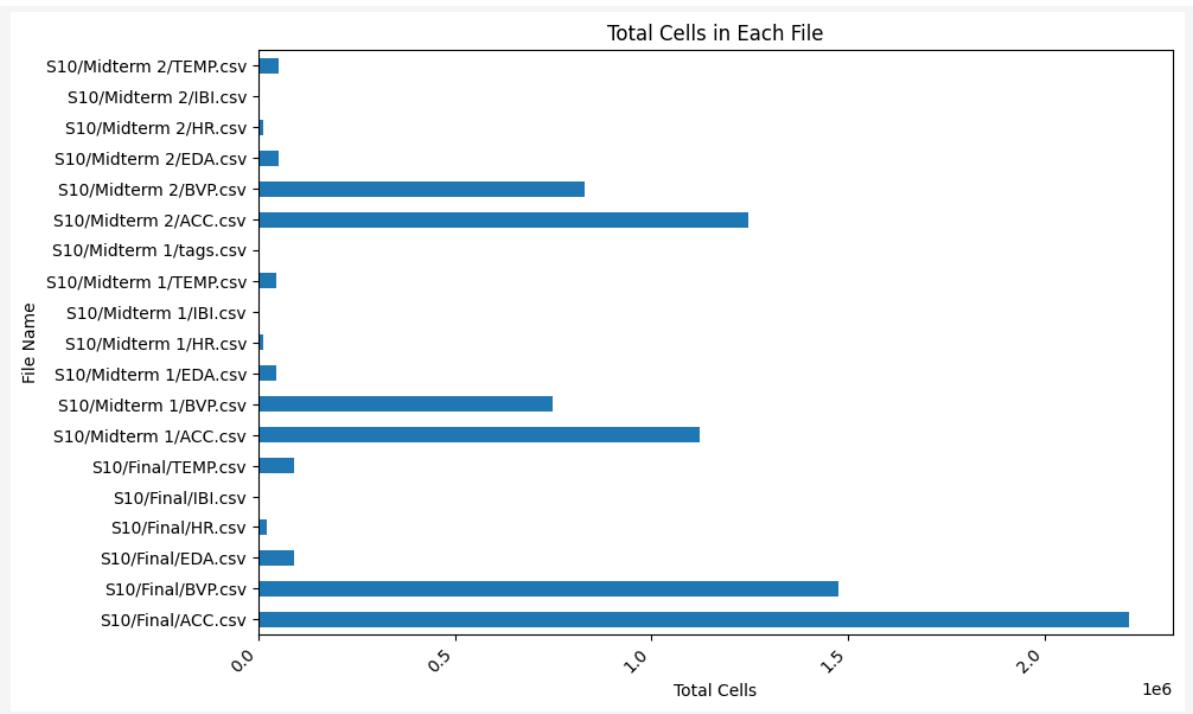
شکل ۶. تعداد سلول‌ها به تفکیک دانشجویان

همانطور که پیشتر توضیح دادم، به کمک تصویر بالا می‌توان به راحتی متوجه شد که تعداد دادهای دیتاست‌ها برای دانشجویان متفاوت است.

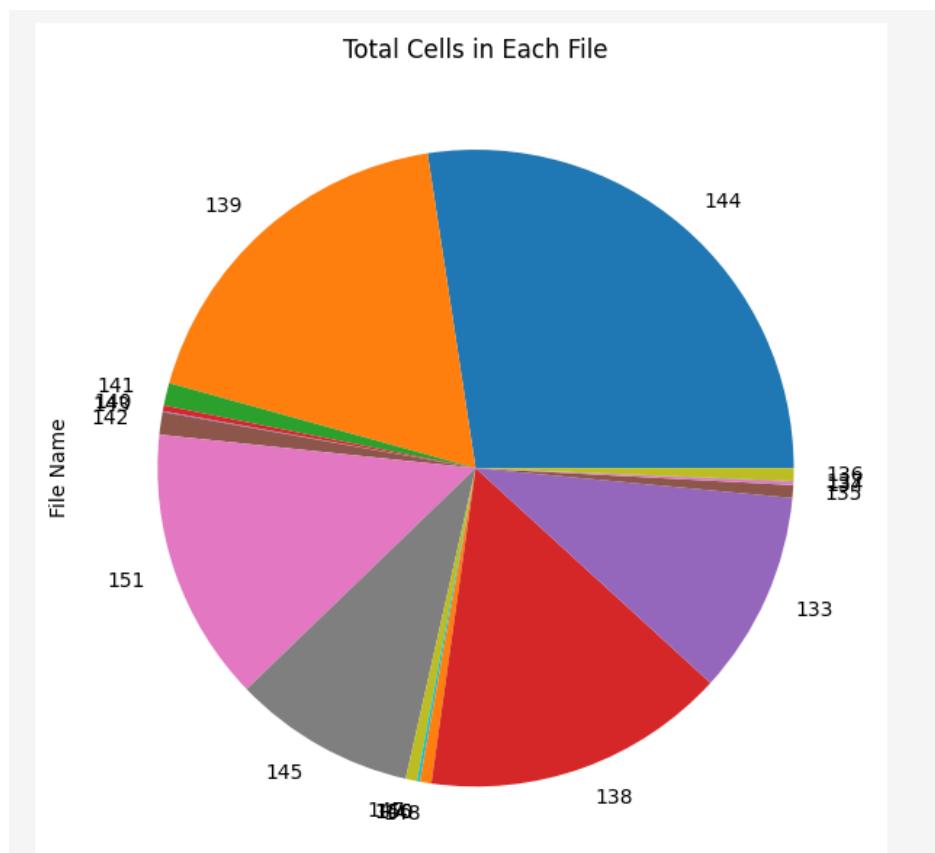
حالا که تمامی دیتاست‌ها را در یک Dataframe قرار دادیم به سادگی می‌توان با نوشتن filter های مناسب خروجی زیر را گرفت.

	<b>File Name</b>	<b>Rows</b>	<b>Columns</b>	<b>Total Cells</b>
125	S1/Final/ACC.csv	748687	3	2246061
120	S1/Final/BVP.csv	1497376	1	1497376
122	S1/Final/EDA.csv	93583	1	93583
121	S1/Final/HR.csv	23388	1	23388
124	S1/Final/IBI.csv	2168	2	4336
123	S1/Final/TEMP.csv	93585	1	93585
132	S1/Midterm 1/ACC.csv	357703	3	1073109
126	S1/Midterm 1/BVP.csv	715408	1	715408
128	S1/Midterm 1/EDA.csv	44713	1	44713
127	S1/Midterm 1/HR.csv	11170	1	11170
131	S1/Midterm 1/IBI.csv	300	2	600
129	S1/Midterm 1/TEMP.csv	44713	1	44713
130	S1/Midterm 1/tags.csv	1	1	1
119	S1/Midterm 2/ACC.csv	356377	3	1069131
114	S1/Midterm 2/BVP.csv	712746	1	712746
116	S1/Midterm 2/EDA.csv	44545	1	44545
115	S1/Midterm 2/HR.csv	11128	1	11128
118	S1/Midterm 2/IBI.csv	1290	2	2580
117	S1/Midterm 2/TEMP.csv	44545	1	44545

شکل ۷. اطلاعات تمامی دیتاستها برای دانشجوی S1



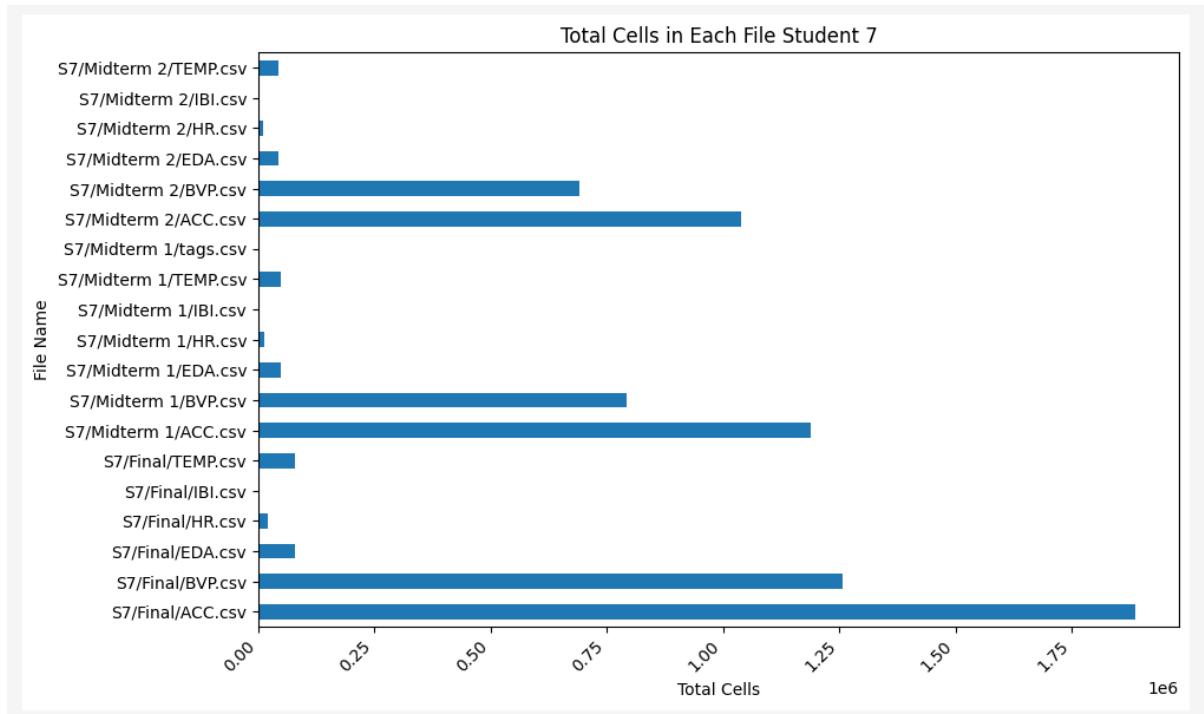
شکل ۸. نمودار میله‌ای برای تعداد سلول‌ها در تمامی دیتاستها برای دانشجو ۵۱۰



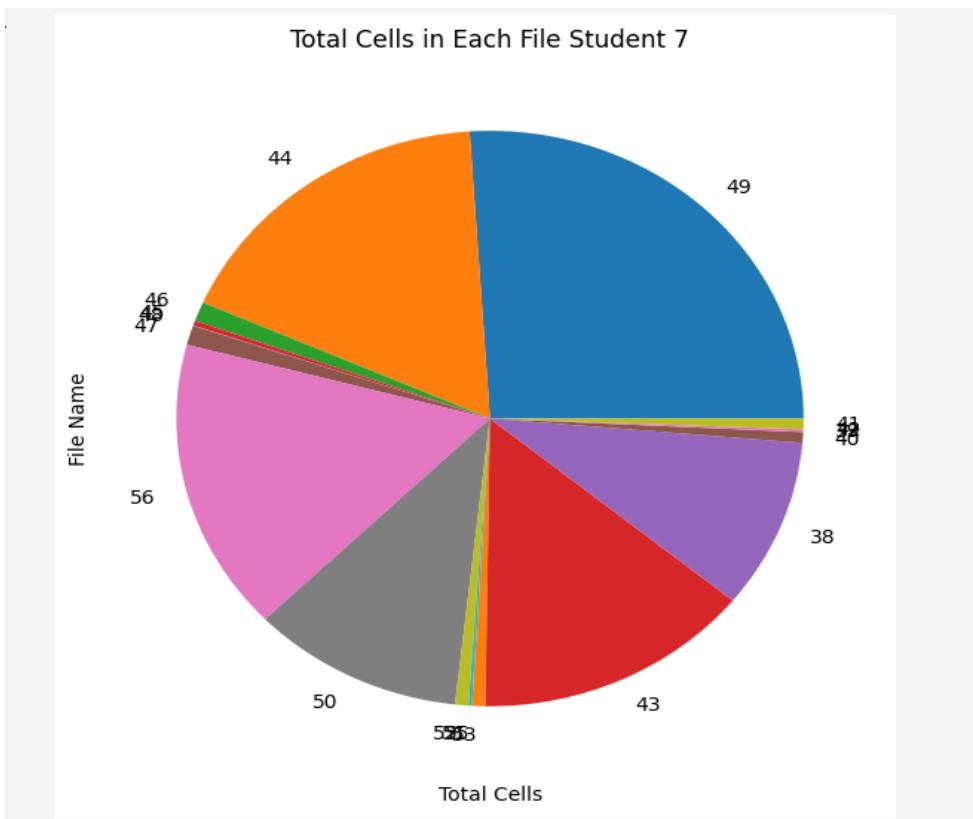
شکل ۹ نمودار pie برای تعداد سلوول‌ها در تمامی دیتاست‌ها برای دانشجویان

به کمک pie چارت بالا می‌توان دید که دیتاست با رنگ آبی که id آن 144 است تعداد بیشتری داده در خود جای داده است و تعدادی از دیتاستها تفاوت بسیار زیادی دارند، مثلاً IBI و HR تعداد بسیار کمتری دارند.

نمودار های زیر برای دانشجوی 7s است.



شکل ۱۰ نمودار میله‌ای برای تعداد سلول‌ها در تمامی دیتاستها برای دانشجو ۷s



شکل ۱ انمودار pie برای تعداد سلول‌ها در تمامی دیتاست‌ها برای دانشجو ۷

با مقایسه این دو دانشجو می‌توان دریافت که این اختلاف واقعاً ممکن است به خاطر ماهیت و جنس داده‌ها باشد چرا که دیتاست‌های مشابه تعدادی نزدیک به عنوان مثال برای دانشجوی ۷ نیز همانند دانشجوی ۱ میزان تعداد سلول‌های دیتاستی چون HR کمتر است.

## اطلاعات کلی

	Rows	Columns	Total Cells
count	1.910000e+02	191.000000	1.910000e+02
mean	2.439213e+05	1.471204	3.926884e+05
std	3.691959e+05	0.752510	5.940370e+05
min	0.000000e+00	1.000000	0.000000e+00
25%	1.142450e+04	1.000000	1.168200e+04
50%	4.884900e+04	1.000000	4.884900e+04
75%	3.876700e+05	2.000000	7.753360e+05
max	1.652608e+06	3.000000	2.478927e+06

شکل ۲ اطلاعات کلی

این جدول آماری اطلاعاتی درباره ستون‌های Total Cells، Columns و Rows ارائه DataFrame summary\_df در است:

### ستون‌ها

- .Rows: تعداد ردیف‌های موجود در هر فایل CSV .1
- .Columns: تعداد ستون‌های موجود در هر فایل CSV .2
- .Total Cells: تعداد کل سلول‌ها در هر فایل CSV (حاصل ضرب Rows و Columns) .3

### توضیحات آماری

- count: تعداد مقادیر غیر تهی در هر ستون.
  - برای هر سه ستون، مقدار 191 است، به این معنی که هیچ مقداری گم نشده است.
- mean: میانگین مقادیر هر ستون.
  - میانگین تعداد ردیف‌ها (Rows) برابر با 243,921.3 است.
  - میانگین تعداد ستون‌ها (Columns) برابر با 1.47 است.
  - میانگین تعداد کل سلول‌ها (Total Cells) برابر با 392,688.4 است.
- std: انحراف معیار مقادیر هر ستون.
  - انحراف معیار تعداد ردیف‌ها بسیار بالا است (369,195.9)، که نشان‌دهنده پراکندگی زیاد در تعداد ردیف‌ها بین فایل‌ها است.
  - انحراف معیار تعداد ستون‌ها پایین است (0.75)، که نشان‌دهنده تعداد ستون‌ها در اکثر فایل‌ها مشابه است.
  - انحراف معیار تعداد کل سلول‌ها نیز بالا است (594,037.0)، که نشان‌دهنده تفاوت زیاد در اندازه فایل‌ها است.
- min: کمترین مقدار در هر ستون.
  - کمترین تعداد ردیف‌ها 0 است (احتمالاً فایل خالی).
  - کمترین تعداد ستون‌ها 1 است.
  - کمترین تعداد کل سلول‌ها 0 است (برای فایل‌های خالی).
- 25% (ربع اول): مقداری که 25 درصد داده‌ها کمتر از آن هستند.
  - 25 درصد فایل‌ها کمتر از 11,424.5 ردیف دارند.
  - 25 درصد فایل‌ها فقط 1 ستون دارند.
  - 25 درصد فایل‌ها کمتر از 11,682 سلول دارند.
- 50% (میانه): مقداری که 50 درصد داده‌ها کمتر از آن هستند.

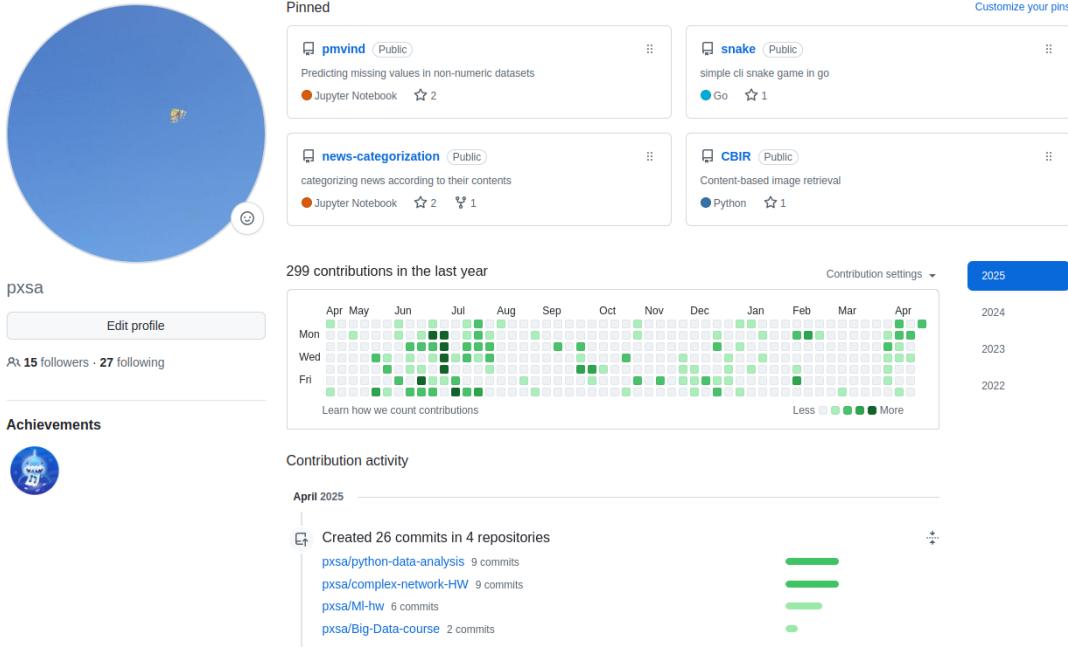
- 50 درصد فایل‌ها کمتر از 48,849 ردیف دارند.
- 50 درصد فایل‌ها فقط 1 ستون دارند.
- 50 درصد فایل‌ها کمتر از 48,849 سلول دارند.
- 75% (ربع سوم): مقداری که 75 درصد داده‌ها کمتر از آن هستند.
- 75 درصد فایل‌ها کمتر از 387,670 ردیف دارند.
- 75 درصد فایل‌ها 2 ستون دارند.
- 75 درصد فایل‌ها کمتر از 775,336 سلول دارند.
- max: بیشترین مقدار در هر ستون.
- بیشترین تعداد ردیف‌ها 1,652,608 است.
- بیشترین تعداد ستون‌ها 3 است.
- بیشترین تعداد کل سلول‌ها 2,478,927 است.

## تحلیل

- پراکندگی زیاد در تعداد ردیف‌ها و سلول‌ها: انحراف معیار بالا و فاصله زیاد بین مقادیر حداقل و حداکثر نشان می‌دهد که اندازه فایل‌ها بسیار متنوع است.
- تعداد ستون‌ها محدود است: اکثر فایل‌ها فقط 1 یا 2 ستون دارند، و حداکثر تعداد ستون‌ها 3 است.
- وجود فایل‌های خالی: مقدار صفر در Rows و Total Cells نشان‌دهنده وجود فایل‌های خالی است.

## گیتهاب

تمامی کدها را می‌توان در آدرس [گیتهاب](#) مشاهده نمود.



شکل ۱۴: گیتھاب

# پایان