به نام خدا

امیرحسین راعقی (۴۰۰۵۲۲۳۷۳)

پروژه Data Science

موضوع پروژه: تحلیل و پیش‌بینی استرس افراد با استفاده از داده‌های چندبعدی

تعریف پروژه:

این پروژه با هدف شناسایی الگوهای استرس در افراد طراحی شده است. استرس یکی از عوامل مهم تاثیرگذار بر سلامت روان و جسم انسان است که می‌تواند بر کیفیت زندگی، عملکرد شغلی و .... تأثیر بگذارد. با بهره‌گیری از داده‌های مرتبط با ویژگی‌های شخصیتی، رفتارهای روزمره (مانند خواب و استفاده از موبایل)، و سنجش‌های فیزیولوژیکی (مانند رسانایی پوست)، هدف این است که مدلی دقیق برای تشخیص سطح استرس ایجاد شود.

اهداف پروژه:

- **شناسایی الگوهای رفتاری و فیزیولوژیکی مرتبط با استرس:**
  تحلیل داده‌ها برای کشف الگوها و روندهای مرتبط با استرس، از جمله نحوه تأثیر ویژگی‌های شخصیتی، عادات خواب، و فعالیت‌های روزانه بر سطوح استرس افراد.
- **تعیین عوامل کلیدی موثر بر استرس:**
  شناسایی مهم‌ترین ویژگی‌های داده که بیشترین ارتباط را با استرس دارند (مانند روان‌رنجوری(تمایل به اضطراب، افسردگی، شک به خود و سایر احساسات منفی)، مدت خواب، یا زمان استفاده از تلفن همراه).
- **درک جامع از ساختار داده‌ها:**
  شناسایی مشخصات کلی داده‌ها از جمله توزیع متغیرها، مقادیر پرت (Outliers)، الگوهای گم‌شده (Missing Data)، و خلاصه آماری هر متغیر.

- **بصری‌سازی داده‌ها برای ارائه بینش:**
  ارائه بصری‌سازی‌هایی از داده‌ها برای ساده‌سازی درک روابط و الگوها، مانند نمودارهای همبستگی، توزیع داده‌ها، و نمودارهای پراکندگی.

کاربردهای پروژه:

- **سلامت روانی:**
  طراحی ابزارهایی برای کمک به افراد در مدیریت استرس.
- **مشاوره:**
  ارائه بینش به روانشناسان و مشاوران برای شخصی‌سازی راهکارهای درمانی.
- **تحقیقات سلامت:**
  کمک به پژوهشگران در فهم بهتر تاثیر رفتارها و عوامل شخصیتی بر استرس.

دامنه,فرمت,حجم,خصوصیات و .... داده‌های پروژه:

پروژه شامل بخش های زیر می باشد:

- **دامنه داده ها:**

شامل چندین بخش متفاوت که در صورت دسته بندی شامل:

1. **ویژگی‌های روان‌شناختی:** شامل صفات شخصیتی اصلی (گشودگی، وظیفه‌شناسی، برون‌گرایی، موافق‌بودن، روان‌رنجوری).
2. **رفتار خواب:** شامل زمان خواب، زمان بیدار شدن، مدت خواب، و کیفیت خواب (PSQI).
3. **فعالیت‌های دیجیتال:** مانند مدت تماس، تعداد تماس‌ها، تعداد پیامک‌ها و زمان روشن بودن صفحه موبایل.
4. **داده‌های فیزیولوژیکی و حرکتی:** مانند رسانایی پوست، فعالیت شتاب‌سنج، شعاع و مسافت حرکتی.
5. **سطح استرس:** با استفاده از نمره مقیاس استرس ادراک شده (PSS).

- **حجم و فرمت و خصوصیات داده ها:**

این دیتاست شامل ۳۰۰۰ رکورد یا ردیف می باشد. که هر کدام از رکورد ها شامل ۲۰ ویژگی می باشد.

این ویژگی ها عبارتند از:

1. Participant_id:(نوع داده= INT): شماره شرکت کننده که در این بخش ما دارای ۱۰۰ شرکت کننده می باشیم.
2. day:(نوع داده= INT): شماره روز مورد بررسی که هر شرکت کننده در ۳۰ روز مورد بررسی قرار گرفته اند.
3. .
4. PSS_score (نوع داده = INT):

نمره مقیاس استرس ادراک شده (Perceived Stress Scale) که سطح استرس روانی فرد را اندازه‌گیری می‌کند. مقیاس این نمره بین ۰ تا ۴۰ است، که نمره بالاتر نشان‌دهنده استرس بیشتر می‌باشد.

5. Openness (نوع داده = FLOAT):

سطح گشودگی به تجربیات جدید و خلاقیت، با مقادیری بین ۰ و ۱.

6. Conscientiousness (نوع داده = FLOAT):

میزان وظیفه‌شناسی، انضباط شخصی، و قابلیت اعتماد، با مقادیری بین ۰ و ۱.

7. Extraversion (نوع داده = FLOAT):

میزان برون‌گرایی و اجتماعی بودن فرد، با مقادیری بین ۰ و ۱.

8. Agreeableness (نوع داده = FLOAT):

سطح مهربانی، دلسوزی، و تمایل به همکاری، با مقادیری بین ۰ و ۱.

9. Neuroticism (نوع داده = FLOAT):

سطح روان‌رنجوری یا ناپایداری عاطفی فرد، با مقادیری بین ۰ و ۱.

10. sleep_time (نوع داده = FLOAT):

زمان خوابیدن فرد، بر حسب ساعت (مثلاً ۲۳.۵ برای ساعت ۱۱:۳۰ شب).

11. wake_time (نوع داده = FLOAT):
زمان بیدار شدن فرد، بر حسب ساعت (مثلاً ۷.۰ برای ساعت ۷ صبح).
12. sleep_duration (نوع داده = FLOAT):
مدت خواب فرد، بر حسب ساعت (مثلاً ۸.۰ برای ۸ ساعت خواب).
13. PSQI_score (نوع داده = INT):
نمره شاخص کیفیت خواب (Pittsburgh Sleep Quality Index). نمره پایین‌تر نشان‌دهنده کیفیت خواب بهتر است.
14. call_duration (نوع داده = FLOAT):
مجموع مدت زمان تماس‌های فرد در طول روز، بر حسب دقیقه.
15. num_calls (نوع داده = INT):
تعداد تماس‌های انجام‌شده توسط فرد در طول روز.
16. num_sms (نوع داده = INT):
تعداد پیامک‌های ارسال‌شده توسط فرد در طول روز.
17. screen_on_time (نوع داده = FLOAT):
مدت زمان روشن بودن صفحه موبایل فرد، بر حسب دقیقه.
18. skin_conductance (نوع داده = FLOAT):
رسانایی الکتریکی پوست، که به‌عنوان یک شاخص فیزیولوژیکی برای استرس اندازه‌گیری می‌شود.
19. accelerometer (نوع داده = FLOAT):
سطح فعالیت ثبت‌شده توسط شتاب‌سنج، که نشان‌دهنده میزان فعالیت بدنی فرد است.
20. mobility_radius (نوع داده = FLOAT):
شعاع حرکتی فرد در طول روز، که نشان‌دهنده گستره حرکات فیزیکی او است.
21. mobility_distance (نوع داده = FLOAT):
مجموع مسافت حرکت کرده توسط فرد در طول روز، بر حسب کیلومتر.

تحلیل کد پیاده سازی شده برای پروژه:

1. در بخش اول کتابخانه های مرتبط با پروژه را وارد می کنیم.

```
• Imports
```

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

2. در ادامه مجموعه داده ها را با استفاده از پکیج **pandas** خوانده و سپس پنج ردیف اول آن را نمایش می دهیم.

```
df = pd.read_csv('stress_detection.csv')
df.head(5)
```

3. سپس نمایش نوع داده‌ها و مقادیر گم‌شده در دیتاست در صورت وجود داشتن:

```
df.info()
```

```
df.isnull().sum()
```

4. نمایش اطلاعات مرتبط با شکل داده‌ها و نام ستون‌ها:

```
df.shape
```

```
df.columns
```

5. خلاصه‌ای از آمار توصیفی مربوط به متغیرهای عددی(میانگین,بیشترین,میانه و ....):

```
df.describe()
```

6. توزیع مقادیر برای متغیر شماره شرکت کننده دسته‌بندی که هر کدام از شرکت کنندگان ۳۰ روز بار تکرار می شوند :

```
df['participant_id'].value_counts()
```

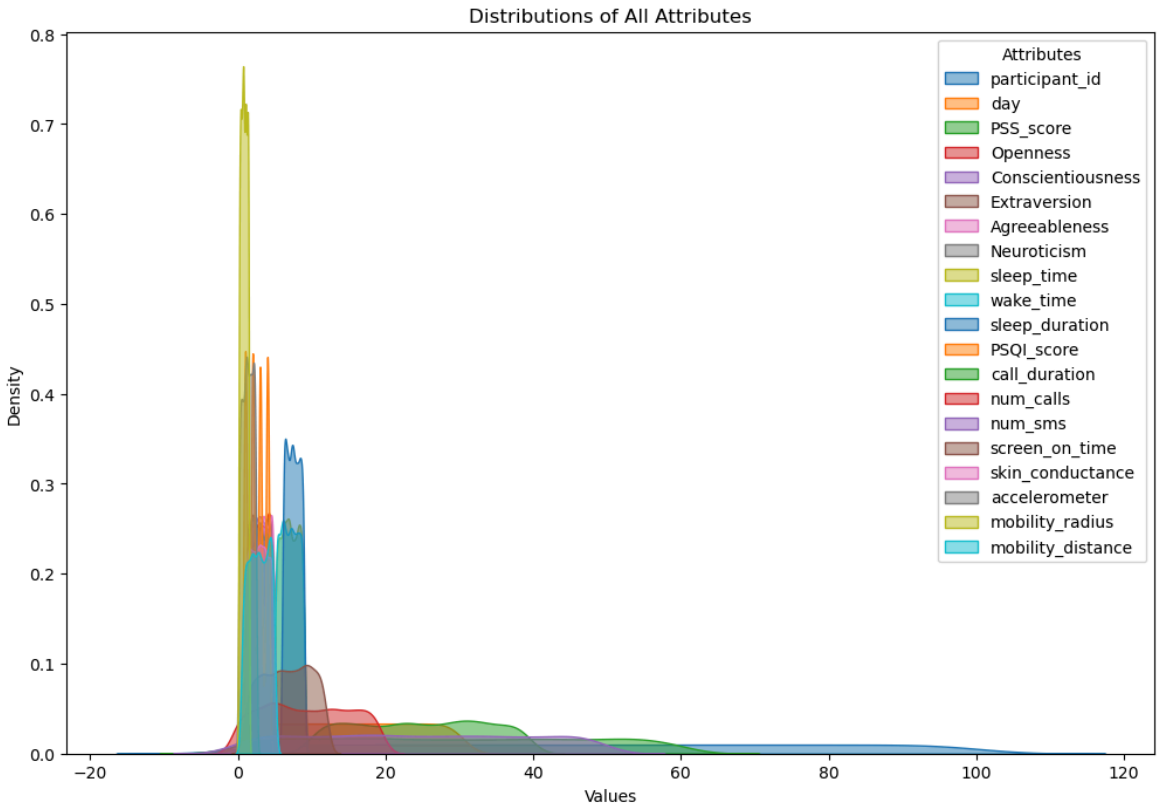
7. در ابتدا بر روی متغیرها به صورت جدا تحلیل انجام می دهیم. در ابتدا توزیع متغیرهای عددی با استفاده از نمودارهای هیستوگرام و KDE را بررسی می کنیم.

کد بخش kde plot:

```
plt.figure(figsize=(12, 8))
for column in df.select_dtypes(include=['number']).columns:
    sns.kdeplot(df[column], label=column, fill=True, alpha=0.5)

plt.title("Distributions of All Attributes")
plt.xlabel("Values")
plt.ylabel("Density")
plt.legend(title="Attributes")
plt.show()
```

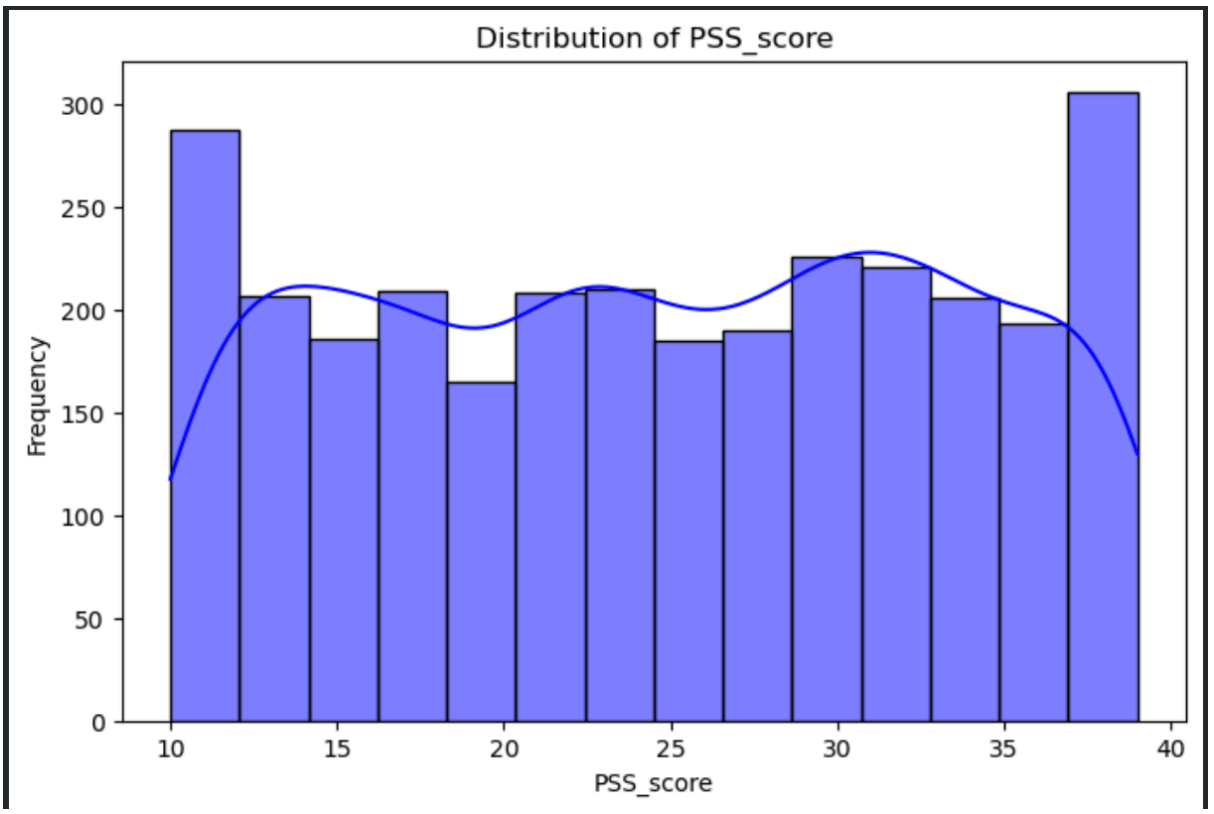
که نمایش آن به صورت زیر می باشد:



کد بخش (Histogram) برای نمایش هر یک از ویژگی ها به صورت جدا:

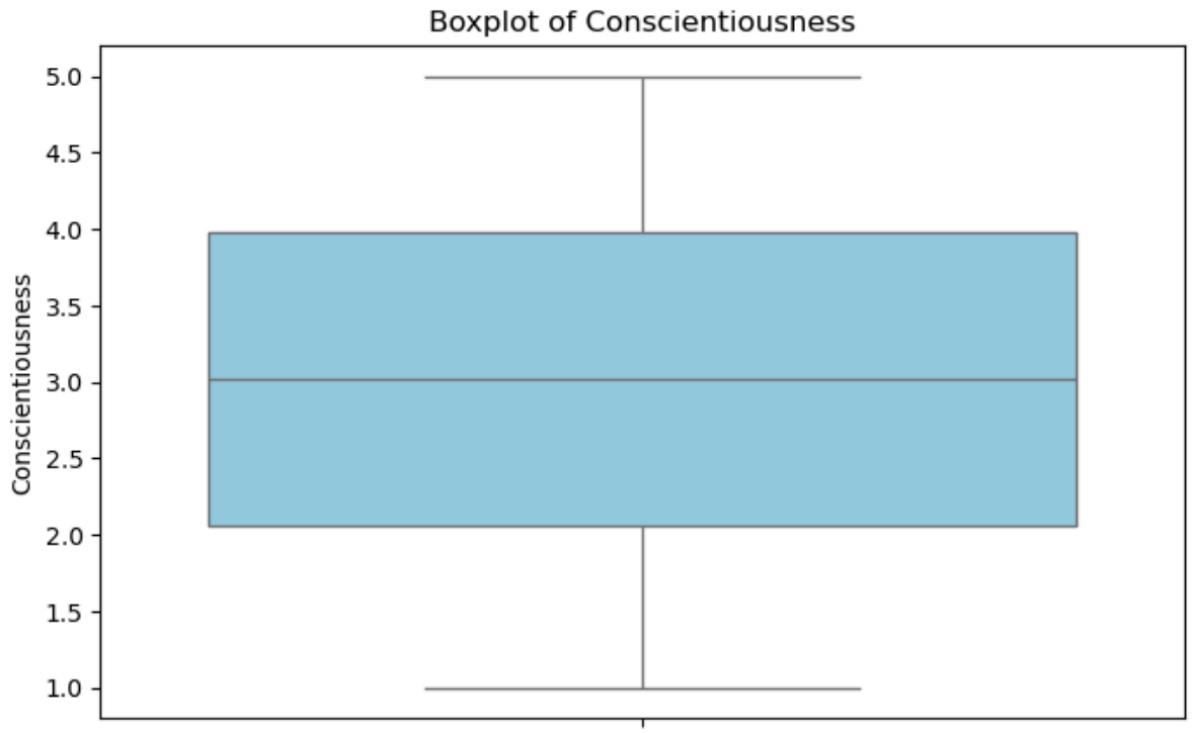
```
for column in df.columns:
    plt.figure(figsize=(8, 5))
    sns.histplot(df[column], kde=True, color='blue')
    plt.title(f"Distribution of {column}")
    plt.xlabel(column)
    plt.ylabel("Frequency")
    plt.show()
```

که نمایش یکی از ویژگی های آن به شکل زیر می باشد:



که طبق شکل بالا توزیع pss_score که در بازه ۰ تا ۴۰ موجود می باشد را نشان می دهد.

در ادامه برای نمایش توزیع پراکندگی هر یک از ویژگی های دیتاست از Box plot استفاده شده است که کد آن به شکل زیر می باشد:

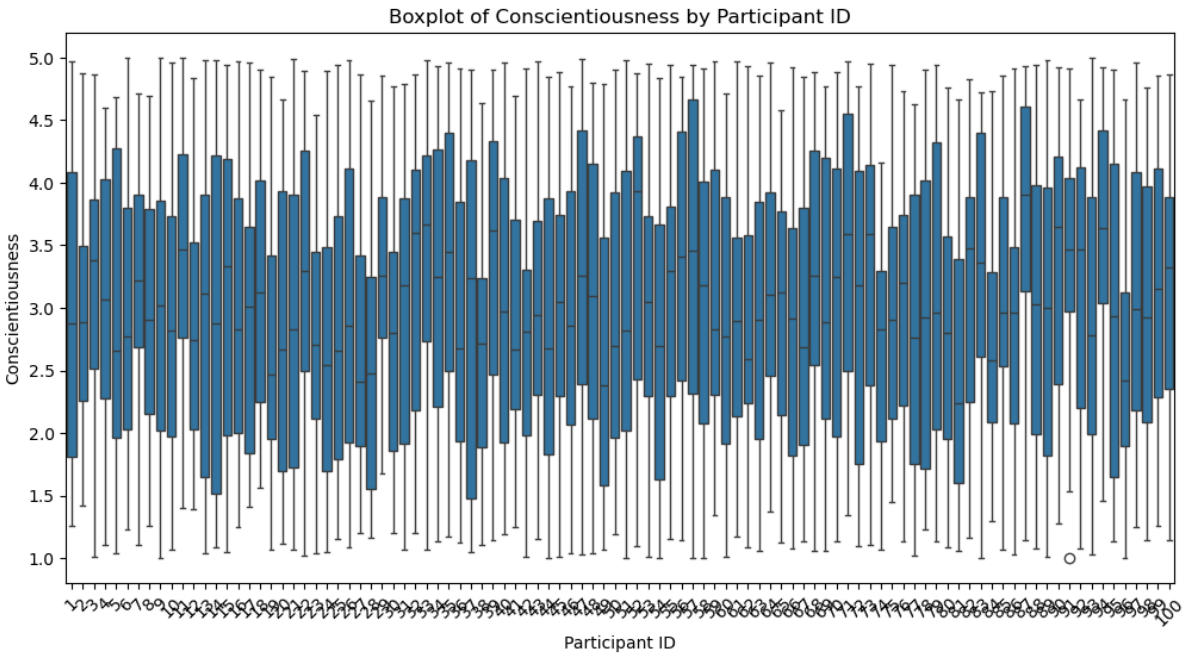


در ادامه برای نمایش دقیق تر شکل Box Plot با کد زیر ویژگی Conscientiousness را برای هر کاربر به صورت جدا ترسیم می کنیم:

```
numeric_columns = Categorical.select_dtypes(include=['number']).columns

for column in numeric_columns:
    if column != 'participant_id':
        plt.figure(figsize=(12, 6))
        sns.boxplot(x='participant_id', y=column, data=df)
        plt.title(f"Boxplot of {column} by Participant ID")
        plt.xlabel("Participant ID")
        plt.ylabel(column)
        plt.xticks(rotation=45)
        plt.show()
```

که نمایش آن به شکل زیر می باشد:



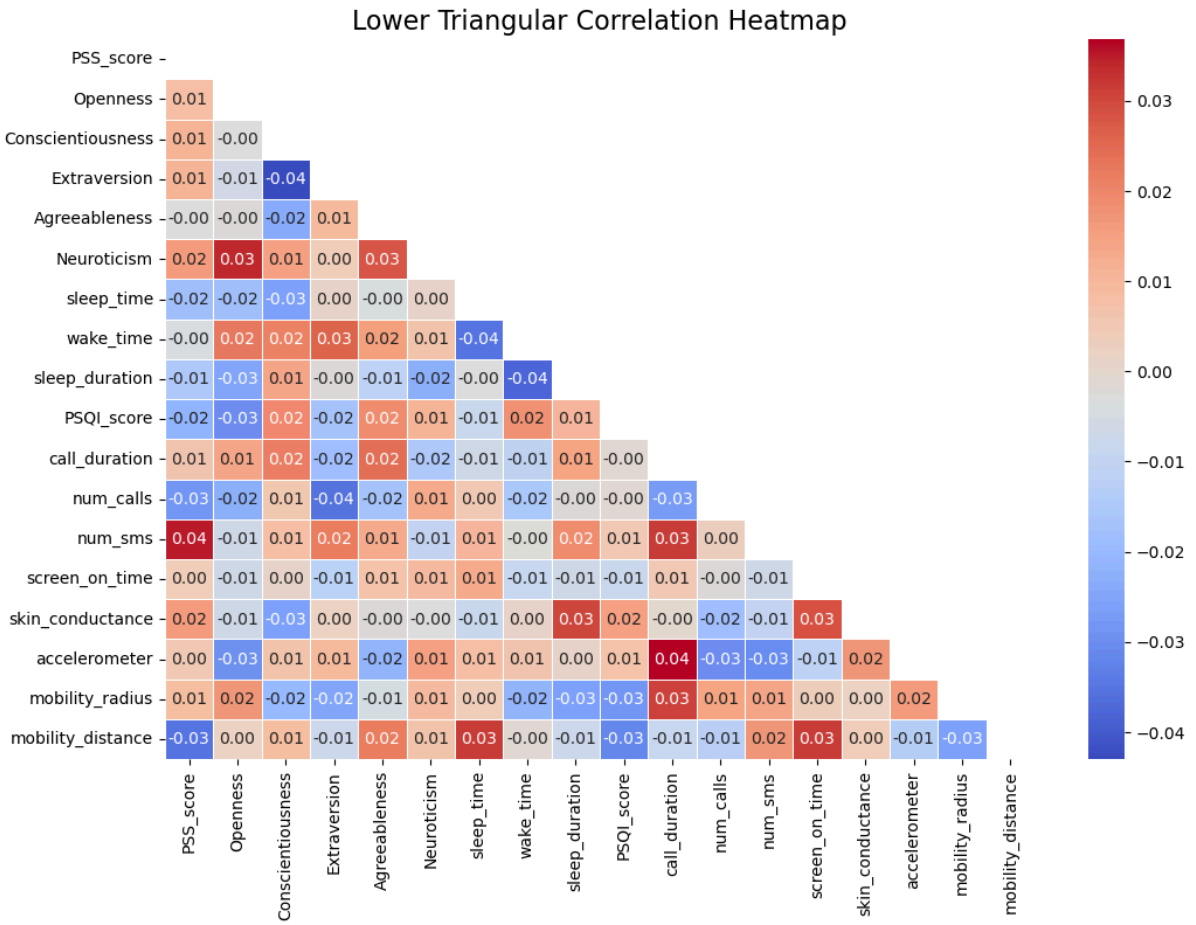
8. حال به سراغ بررسی بر روی چندین متغیر و نمایش ارتباط میان چندین متغیر با هم می پردازیم. ابتدا نمایش Heatmap برای دیتاست فعلی ترسیم می کنیم کد:

```
corr = HeatMap.corr()
mask = np.triu(np.ones_like(corr, dtype=bool))

plt.figure(figsize=(12, 8))

sns.heatmap(corr, mask=mask, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Lower Triangular Correlation Heatmap", fontsize=16)
plt.show()
```

که با توجه به Heatmap تصویر زیر را نتیجه می گیریم:

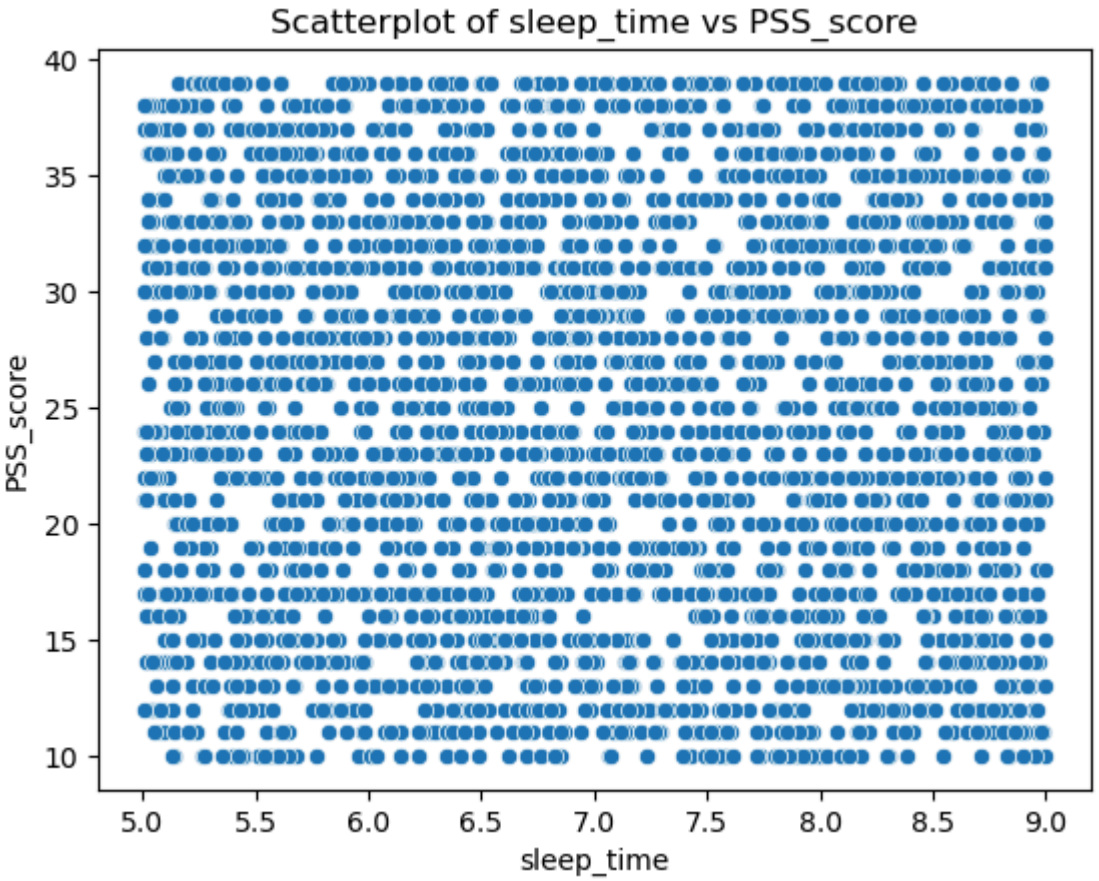


همانطور که در نمایش Heatmap معلوم هست رابطه تک تک ویژگی های دیتاست را با یکدیگر مقایسه می کنیم. در نمودار هرچه ما به ۱ نزدیک تر شویم به معنای آن است که ۲ ویژگی با هم ارتباط مستقیم دارند یعنی با زیاد شدن یکی دیگری نیز افزایش می یابد و هرچه به -۱ نزدیک شوند یعنی دو ویژگی با هم رابطه عکس دارند با اضافه شدن یکی از دیگری کاسته می شود. که اعداد هرچه به رنگ منفی میل کنند آنها را با رنگ آبی و هرچه عدد به مثبت ها میل کند آن را با رنگ قرمز نمایش می دهیم به طور مثال در شکل بالا ما متوجه می شویم که مقدار استرس درک شده توسط شرکت کننده با مقدار پیامک های ارسالی او رابطه مستقیم دارد.

9. نوعی دیگر از نمایش بر روی چندین متغیر و نمایش ارتباط میان چندین متغیر با هم به صورت دو به دو با کمک از Scatter plots می باشد که کد آن به شکل مقابل می باشد:

```
for column in Scatter.columns:
    if column != 'PSS_score':
        sns.scatterplot(x=column, y='PSS_score', data=df)
        plt.title(f"Scatterplot of {column} vs PSS_score")
        plt.xlabel(column)
        plt.ylabel('PSS_score')
        plt.show()
```

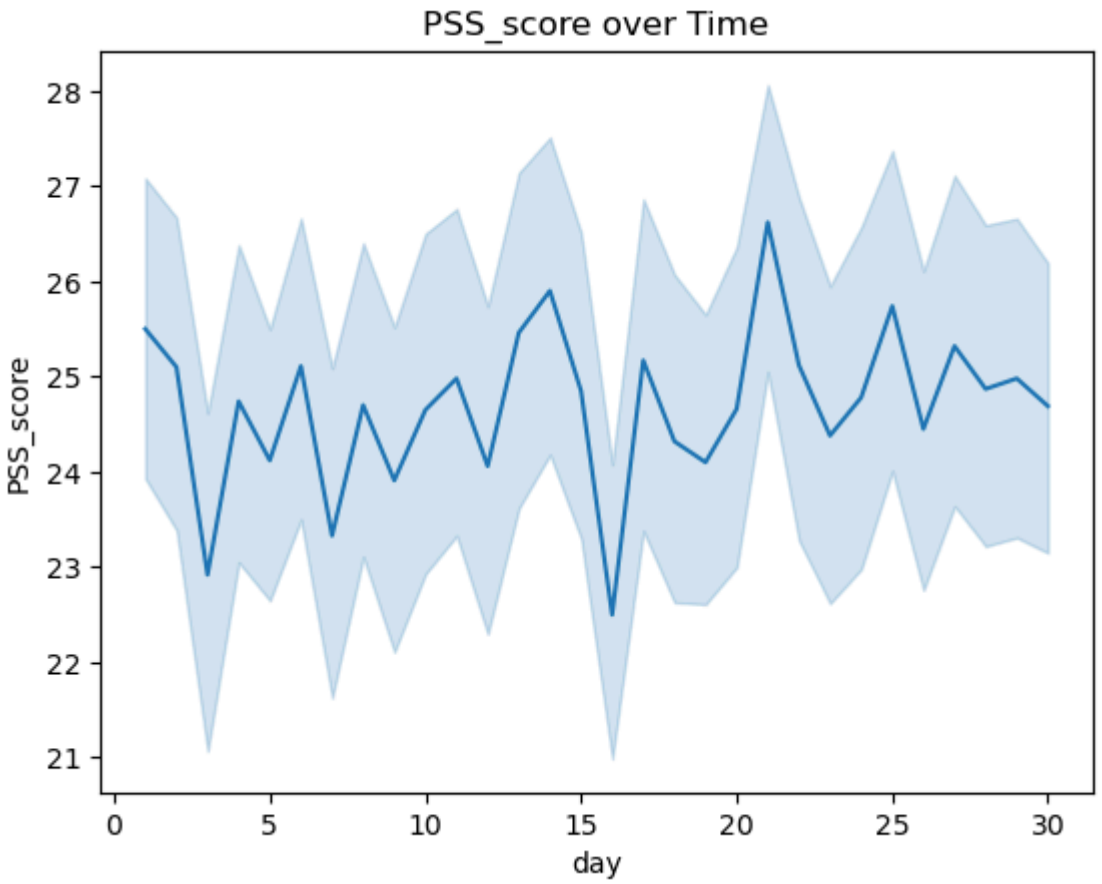
که نمایش آن برای مقدار خواب و میزان استرس به شکل مقابل می شود:



10. در تلاشی دیگر برای پیدا کردن ارتباطی میان سطح استرس و زمان گذشته طی روز های متفاوت توسط کد مقابل:

```
sns.lineplot(x='day', y='PSS_score', data=df)
plt.title('PSS_score over Time')
plt.show()
```

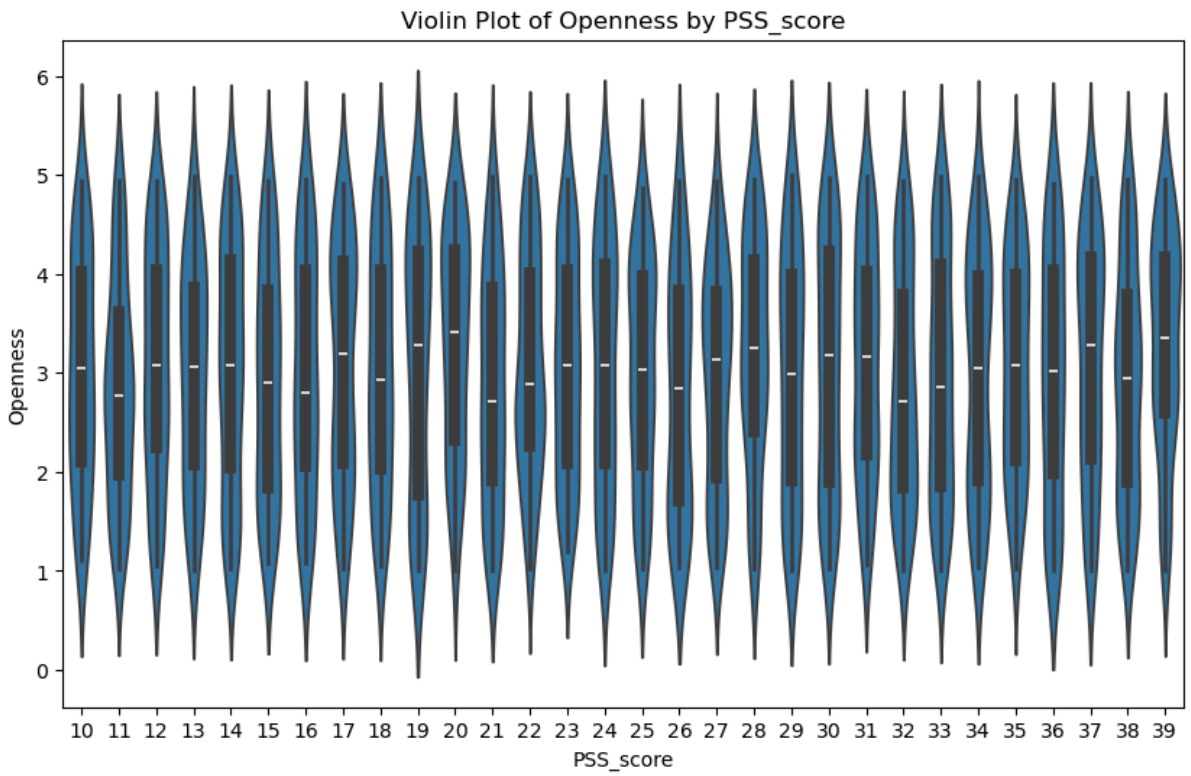
که حاصل متوسط شرکت کنندگان آن به صورت زیر می باشد:



11. نوعی دیگر از نمایش که مورد استفاده قرار گرفت استفاده از violin plot می باشد که از طریق کد زیر به آن می رسیم:

```
for column in Violin.select_dtypes(include=['float64', 'int64']).columns:
    if column != 'PSS_score':
        plt.figure(figsize=(10, 6))
        sns.violinplot(x='PSS_score', y=column, data=df)
        plt.title(f'Violin Plot of {column} by PSS_score')
        plt.xlabel('PSS_score')
        plt.ylabel(column)
        plt.show()
```

که حاصل آن شکل زیر می شود:



تفاوت این روش نسبت به روش Box plot ارائه اطلاعات بیشتر که شامل چگالی احتمال و ... می باشد معمولا برای زمانی است که مقدار سمپل های ما زیاد نباشد. Box plot به صورت خلاصه ولی در این روش توزیع کامل داده را به ما نمایش می دهد.

12. و در آخر سر بر اساس خوشه بندی به دنبال پیدا کردن ارتباطی میان ویژگی های دیتاست بودم که کد آن به صورت:

```
sns.clustermap(Cluster.corr(), annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Cluster Map of Feature Correlations')
plt.show()
```

و نمایش آن به صورت زیر می باشد:

Cluster Map of Feature Correlations