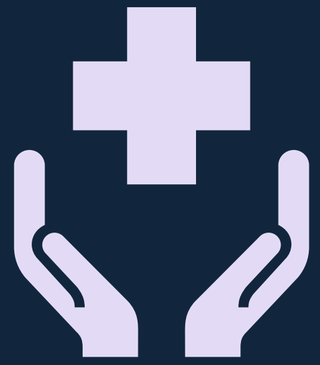
Mental Health & Emotional Stability

Drug Prevention & Rehabilitation Foundation



Nadeen Ali, April 20th 2024

The Fundamentals of Data Analytics & Their Uses

DRUG PREVENTION & REHABILITATION FOUNDATION

Data Analytic Activities, Techniques, and Tools

Greater detail will be discussed in next parts.

Activites

This is the **process** of studying, understanding, and working with the data. Working with the data includes ***stating (defining) the question &/or problem*** to properly understand the scope and goal of work and analysis. Working with data also includes a wide range of steps as the following: ***collection, exploration, cleaning, analyzation, and interpretation***. Once the analysis is over, **result communication** is presented to share findings, understandings, and possible solutions.

Techniques

In short, techniques is the specifics of *what* is done like:

measures of -

- ***central tendency*** (mean, median, mode)
- ***dispersion*** (standard deviation, variance, coefficient of variation, range, IQR)
- ***frequency*** (frequency, plots)
- ***position*** (decile, quartile, percentile, z-score)

And, of course, **visualization & graphs** like bar & pie charts, histogram, density & box plots, heatmap.

Tools

Tools includes the specifics of *what* to use or *how* to do. Tools can be as basic as programming languages like **python** which is widely used in data science & analysis due to its simplicity, libraries, and large open-source community. In order to use a dataset, spreadsheet tools like excel is used. It is read through a library call ***pandas***. Data visualization tools include the libraries like ***matplotlib*** & ***seaborn***. Other libraries like ***sklearn*** [*label encoder*] help turn categorical data into numerical for many uses, as well as ***scipy*** [*z-score*] to calculate the z-score of the data.

Data Analytic Methods: Intro

A Story About The Case Study

- This study was aimed at people who have taken or highly contemplated taking drugs during the corona quarantine period to further study the long lasting effect the pandemic has on drug abuse as a coping mechanism. The study focused on testing a wide range of possible factors from basic, personal information about themselves and lifestyles to daily habits, physical appearances, and mental health. Understanding *suspected* drug takers and their mannerisms both before *and* after abusing drugs is crucial for a hefty number of reasons as this can help us with both detecting early signs of a 'to-come' drug abuser as well as root problems for current drug abusers. Once the factors have been found and studied, I as a member of the Drug Prevention and Rehabilitation Center can put into motion ways for those struggling (abusers and those contemplating) to seek help through us and our future aid.

- Drug usage can lead to social and personal impacts, destroying both the user and potentially their environment, so studying a matter as crucial as this shall not be taken lightly, and is therefore why we have paid great attention to it. In order to reach our goal, a wide range of data analysis methods has been implemented to reach a number of pattern recognition, trends, and decisions. The following points discuss these methods and processes briefly as greater detail will be discussed in later parts:

Data Analytic Methods: Descriptive Analysis

A Story About The Case Study

1. Descriptive Analysis: During this stage, I've explored lots of data regarding mental health and emotional stability and applied many different techniques and tools for a deeper understanding of the study and data I'm working with. I will dive into greater detail in the following parts but, as a brief summary, I have explored various measures for the 'Weighted Stress Index' of the study subjects to learn the range of the stress, the average, and even what factors affect it the most. I have also explored the relationships between the different data topics surrounding mental health and emotional stability in order to find the ratio of yes & no answers for each one, if any factor directly affects another, how they relate to drug abuse, and much more. And, of course, many visualization methods were used in order to properly *see* the data with the eye, allowing for better interpretation of the numbers I'm reading.

Data Analytic Methods: Predictive Analysis

A Story About The Case Study

2. Predictive Analysis: This 2nd stage includes using the past studied data in the 1st stage of analysis (descriptive analysis) to predict possible future events based on the trends, patterns, and relationships. As a quick example, I was able to find that the mood and emotional control of the study subjects greatly affected the outcome of drug usage. With this information, I can take the answers of these 2as the main factors and predict whether a person abuses drugs or not through them. How accurate it is, though, also depends on other factors, but it is a great baseline and starting point. These 2 factors also highly affect the outcome of the weighted stress index, allowing me to predict the stress of a person based on them as well.

Data Analytic Methods: Prescriptive Analysis

A Story About The Case Study

3. Prescriptive Analysis: In short, this is the *last* stage and the decision making stage. Going off the quick predictions I have mentioned in stage 2 of the analysis (*predictive analysis*), if someone shows signs of irritable mood and difficulty controlling emotions, then they should be prioritized to study on possible drug intake. *If* their drug testing results come back as negative, it is still important to maintain close attention to them as this can be a sign of potential drug use in the future. Giving them the help they need to manage their emotions and live peacefully is important to shelter people as so from bad coping mechanisms.

Descriptive Analytic Techniques

Weighted Stress Index: Measures

Measures of CENTRAL TENDENCY

These measures include the mean, median, *and* mode (although mode is oftentimes used for categorical data rather than numerical). The results is a single value (except mode can be multiple) that describes the data's central point in some way. Here's some greater detail:

1 - **Mean = 16.869:** average value of stress using the technique .mean()

2 - **Median = 19.92:** the very middle value in the dataset using the technique .median()

3 - **Mode = 0.1, occurrence = 30.23%:** the most occurred value(s)

While these measures and values are inherently useful, it is quite difficult to make some sense of them without some extra information... such as the measures of range.

Are they ***related??***

Measures of DISPERSION

Now, these measures test the spread of the points or more specifically the distance between them and their central tendency calculated earlier. If dispersion is low, the values are close to one another - if high, the values are more spread out. Here are the ways to calculate:

1. **Range:** Subtracting the min value from max, we can find the general spread of the data
   - value = **~33.79**, max = ~33.79, min = 0.0
2. **Standard Deviation (std):** the average deviation or distance from the mean using the technique .std()
   - value = **~11.73**
3. **Variance:** standard deviation but squared using the technique .var()
   - value = **~137.597**
4. **Coefficient of Variation (CV):** ratio of std to mean, allowing us to see how big or small the std is compared to the mean. also, it allows direct comparisons with CV values of different scales, means, or units. this is found using the technique .cv()
   - value = **~69.536**
5. **Interquartile Range (IQR):** this touches onto measures of *position* which is next slide. however, it is the spread of the *inner,* middle quartile (Q2) by subtracting Q1 from Q3. it's usually unaffected by outliers and gives an idea of the spread of data around the median.
   - value = **~23.323**

Cont. – Insights

What I Found

The average person faces an average amount of stress according to the mean (16.869, ~50% of the max of 33.79). Also, the median shows that 50% of people have a stress level over and under ~19.92, with the lower 50% having a greater range than the higher 50% meaning people are quite averagely to highly stressed. Though mode isn't typically used for *continuous numerical* data, a stress of 0.1 takes up **over 30%**, showing some bias. Lastly, the large IQR (~23.323) shows a wider spread of stress within the middle 50%, meaning there's lots of variability there.

Considering a total of 227 persons were studied, the dataset is quite *small*. Although the ***mean*** (16.869) and ***median*** (~19.92) may allow the assumption that the data is distributed evenly according to the ***range*** (33.79, same as max & min is 0), the measures of *dispersion* tell a different story.

The standard deviation of ~11.73 calls for a different story, showing that the average deviation from the mean is relatively large considering the range. The coefficient of variation (~69.536) proves this further, as a value greater than 1 indicates a huge difference between the std and mean of the data.

What Does it Mean?

This can mean a few things:

1. **Unbiased Dataset:** A great dispersion in the data usually means there is little bias as each person, to some extent, is unique in their own way. This always the study & analysis to, potentially, be more catered towards the population *as a whole* rather than a focus group or a sample.
2. **Lack of Focus:** However, this lack of bias can also mean a lack of *focus* as each study subject is quite different from the other, making it harder to find solid information from easily recognizable trends and patterns.
3. **People are Averagely Stressed:** From the mean, we can tell the average stress is 50% of the max. Though the median suggests people are a *little more* stressed, it's still quite a close number.

Measures of FREQUENCY

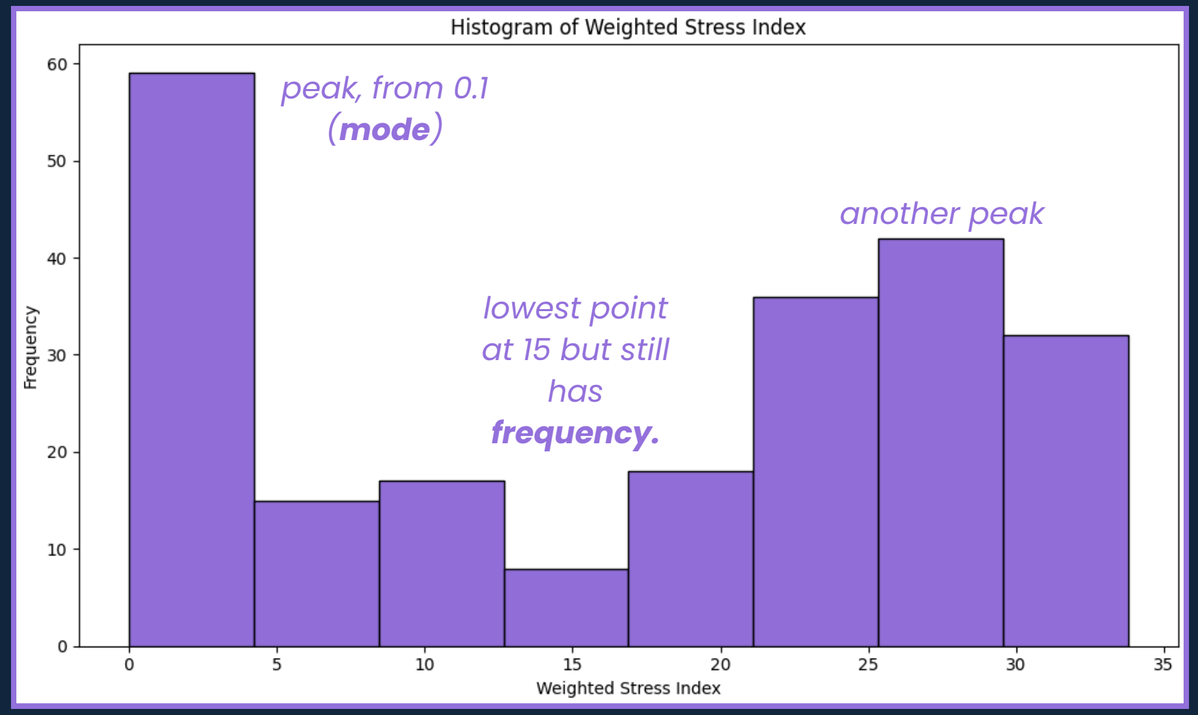
These measures simply calculate the number of times some value occurs in multiple ways. In this particular study, I have only used the base *frequency* and *relative frequency* (will be visited later). Visualization tools like the histogram, bar chart, and pie chart can be used to represent these measures as well.

These measures help us understand what values appear the most often, allowing us to draw many conclusions like potential focus groups, popularity, etc.

Insights

As mentioned earlier about the potential bias from the mode of **0.1** that takes up over **30%** of the stress value, it can be clearly seen in the histogram as its bin from 0 - 4 is the *highest peak*. This is definitely affecting the outcome of other measures and leading to some bias potentially.

Additionally, there seems to be no outliers as every bin, even the lowest frequency, has some frequency to prove its existence. More details on outlier detection will be discussed in the next slides.



HISTORGRAM: Explanation

The *technique* histogram divides continuous values like the stress index here into bins with some value range. It shows the *frequency* or how much a certain range of values is shown. Clearly, the range from 0 to 4 is peak, then the stress values decrease quite a bit and begin increasing again till it reaches another peak, though smaller, at range from 21 to 25. This was done through the matplotlib library as a tool

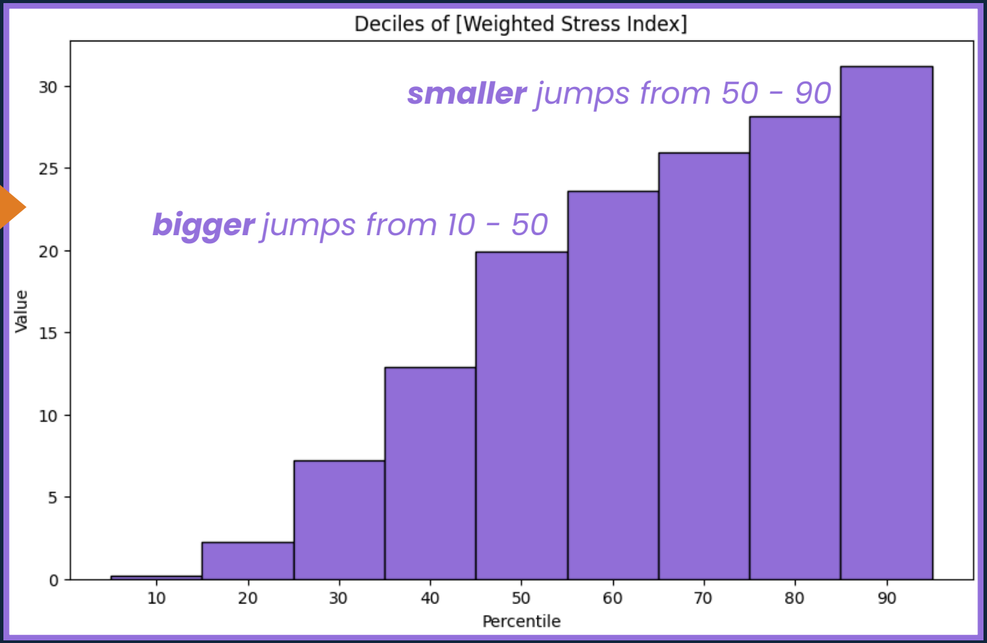
DRUG PREVENTION & REHABILITATION FOUNDATION

Measures of POSITION

These measures, from the name, show the positions of datapoints within the data.

1 - Decile: divides the data into 10 equal parts from 10% (D1) to 90% (D9)

values =
10th Decile: 0.2
20th Decile: 2.25
30th Decile: 7.19
40th Decile: 12.897
50th Decile: 19.921
60th Decile: 23.607
70th Decile: 25.928
80th Decile: 28.12
90th Decile: 31.192

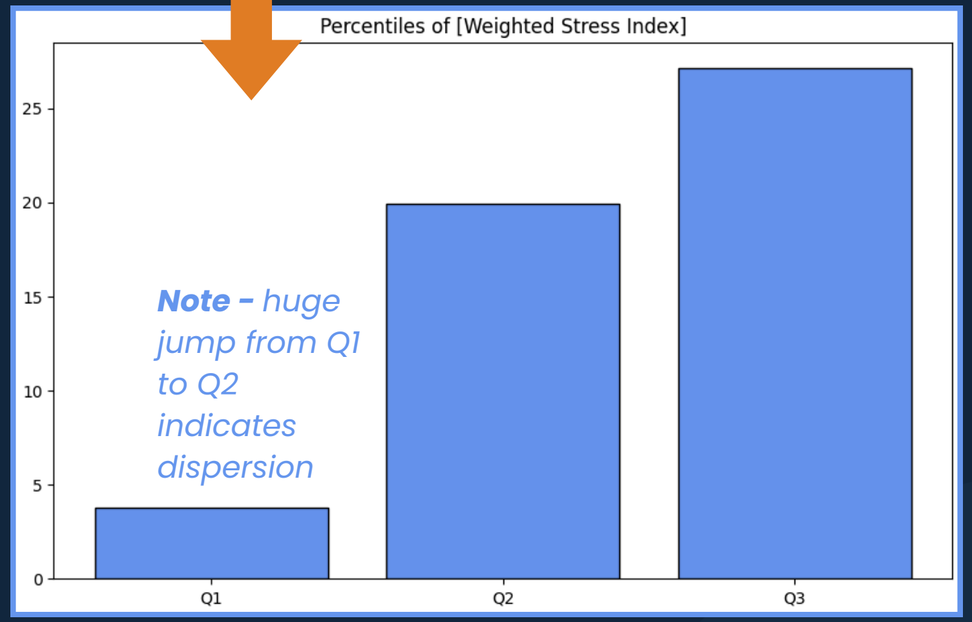


Note - *The usage of these* ***bar charts*** *allows the eyes to envision how the data jumps from one percentile to another*

Tools & Techniques - To calculate the percentiles for both decile & quartile, numpy was used as a tool to gain access to the percentile technique. Matplotlib was then used as a tool to use the technique of bar chart for the visualization of the data.

2 - Quartile: divides the data into 4 equal parts from 25% (Q1) to Q3 (75%)

values =
Q1: 3.8
Q2: 19.921
Q3: 27.123

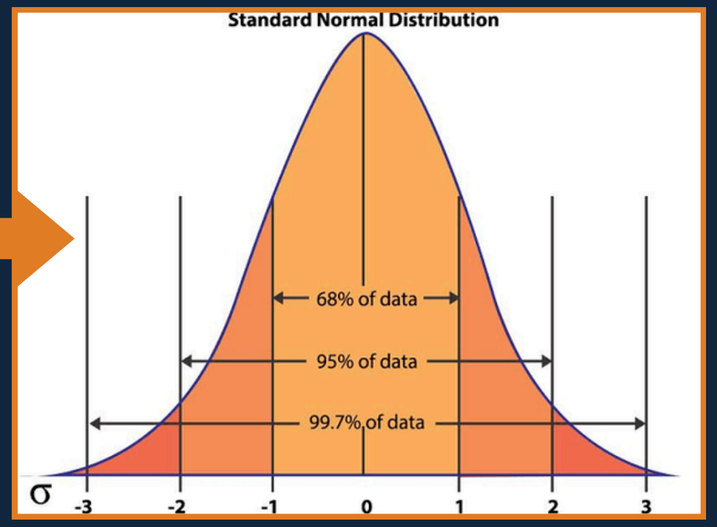


Measures of POSITION

Percentiles alone cannot really be understood deeply, therefore the following is *also* used in conjunction:

3 - Z-Score: the number of standard deviations a point is *away* from the mean. Using zscore & stats from scipy.stats, I got the following Z-scores:

values =
*M*ax: 1.446
Min: -1.441
Range: 2.887



Note - *As we can see in the Normal Distribution graph, all of the data is contained with no outliers.*

Tools & Techniques -
To calculate the z-score, library scipy was used as a tool to get access to the zscore & stats techniques

4 - Outlier Detection(s): Using 4 different methods, it is *proven* that the data within the stress index column does *not* have any outliers. These methods are:

- IQR
- Z-Scores
- Box Plot
- Density Ploto

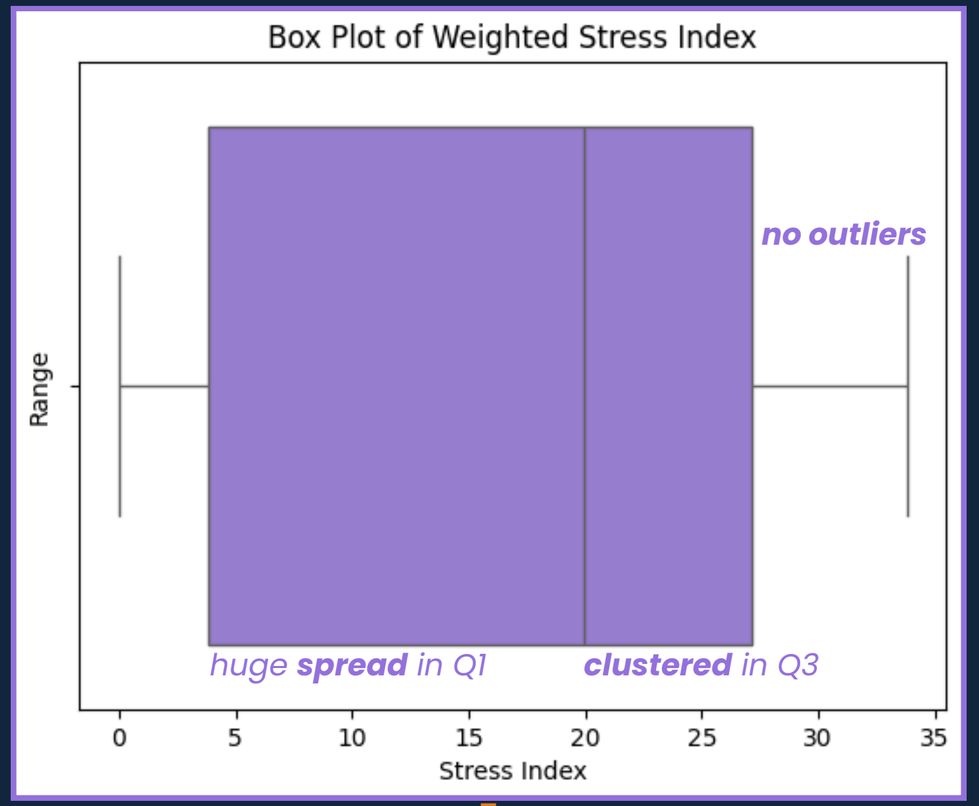
and ***how's it done??***

```
Lower Bound Outlier(s):

Empty DataFrame
Columns: [Case_ID, Age,
Index: []

[0 rows x 60 columns]

----------

Higher Bound Outlier(s)

Empty DataFrame
Columns: [Case_ID, Age,
Index: []
```

df size when ***extracting outliers***

IQR Method: By calculating the min (Q1 - 1.5 * IQR) & max (Q3 + 1.5 * IQR) scores, I've found that no stress values are *under* the min or *over* the max.

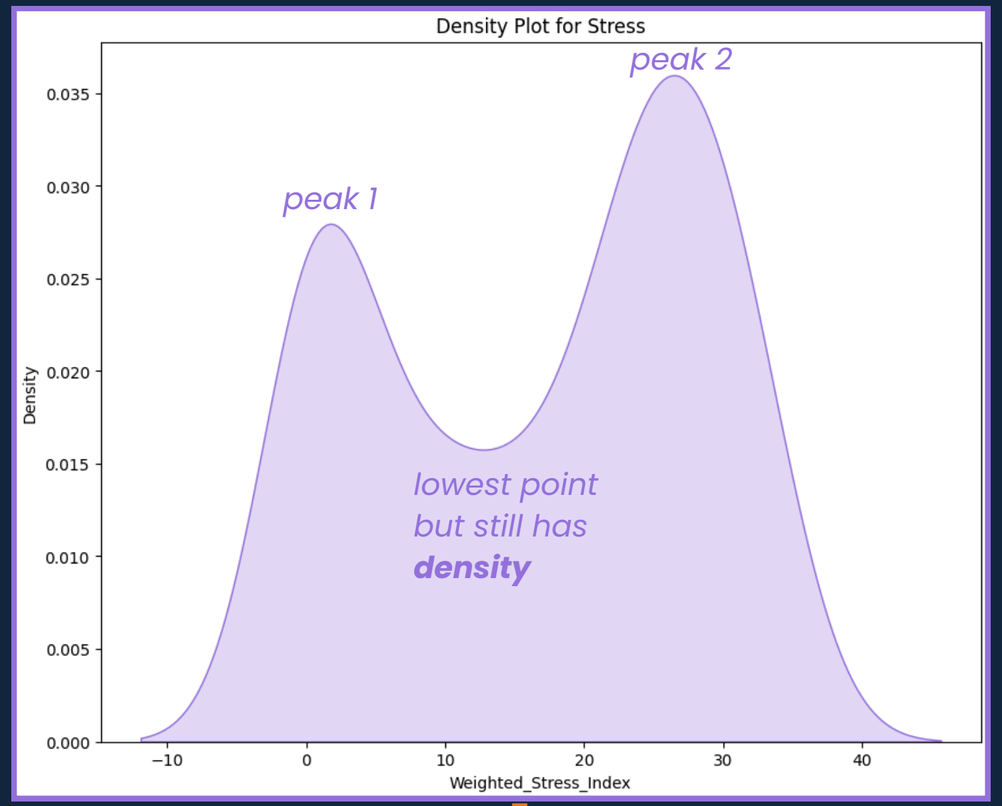
Z-Score Method: Similarly, as seen in the graph, I've looked for the values *under* -3 and *over* +3 for outliers, but there were **none.**



Measures of POSITION

And here's a continuation of the *plots* of position and outlier detection.

The techniques of the box & density plots were created through the libraries or tools of seaborn and matplotlib



BOX PLOT: Explanation

As we can see in the box plot above, there are no circles *outside* the min & max score lines, indicating no stress values fall outside into the outlier regions.

DENSITY PLOT: Explanation

Similarly, the density plot (a smooth histogram pretty much), shows the concentration of the continuous values of the stress index. Despite highs & lows, none of the intervals ever go so low or almost non-existent, indicating no outliers.

Cont. - Insights

What I Found

To recap, I've found that there's quite the big jumps in values from the D1 to D5 (0 to 20), and smaller jumps from the D5 TO D9 (20 to 32). The deciles also show the great range of stress levels as it goes from 0 to 32.

This is further proved in bar chart for the quantiles, showing a *huge* jump from Q1 to the Q2 (4 to 20), but a much smaller difference between Q2 and Q3 (20 to ~27). This means the data begins to spread greatly during the first 50% and the difference begins to slow down towards the end of the 50%.

In addition, as I've already gone through, I've come to find there are *no* outliers within the data of Weighted Stress Index. Using the 4 methods of percentile, z-score, box plot, *and* density plot, all of the data is seen to be included in the non-outlier range. All data was within a z-score range of -1.441 to 1.446, so extreme stress levels are not found.

What Does it Mean?

This can mean a few things:

1. **Unbiased Dataset:** Data without *any* outliers is often a sign of unbiased data as no specific value *diverges* the data. Outliers can and will affect the mean, standard deviation, variance, etc.
2. **Overfitting:** *However*, in some cases, having outliers may be a good thing. In my case working in the Drug Prevention and Rehab Foundation, it is important to have information on every possible & realistic set of answers and test results to generalize the population and derive accurate insights. This also allows me to find solutions for a greater number of people to benefit from rather than a certain (& accidental) focus group.
3. **Most People are Averagely - Highly Stressed:** According to the deciles *and* quartiles, we can see that the 2nd half of the study subjects are clustered between a level index of around 20 to 30 compared to a spread from 0 to 20.

DECISIONS

Therapy Sessions & Workshops for Highly Stressed Individuals -
There are small differences between the high stress values between Q2 and Q3 or D5 to D9. These sessions & workshops will focus on further understanding the patients' lives, habits, emotions, etc. to solve their root problems.

Aiming for a Stress Level Within Q1, D1 to D3, *or* Mean -
This data can be used to set a realistic goal for the patients to reach the average stress level.

Communicational Exchange with Individuals in Stress Levels D1 to D3 -
A group of the lesser stressed people can exchange stories, experience, and advice to aid the higher stressed people with improving through talking.

Focus Group with Individuals of 0.1 Stress Value -
Further study and understand the behavior of these incredibly low and dominant stress index value (found from mode and histogram). Perhaps they can also join the exchange discussion to further aid the stressed people.

How Did **Descriptive Analytics** Help?

Everyone of the descriptive analysis methods I used served an important cause and help. Here are some greater details:

Mean, Median, Mode - These allowed me to gain a general idea of the data. Using the average stress, for example, the mean can allow me to understand if someone's stress is high or not. Or, I can also use the median to determine in which half the person's stress lies in. The mode aids with potential division to find *focus groups* - the value 0.1 was highly repeated, therefore I can find a way to take advantage of this group of people.

Range (with max, min) - This gives a basic look on the spread of data to

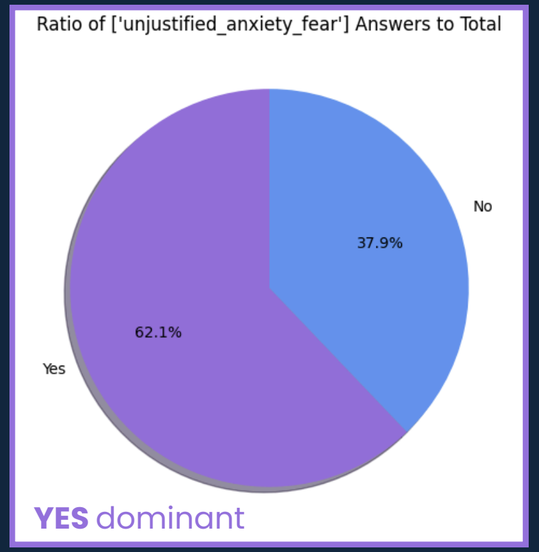
Categorical Data:
Measures of Frequency

Measures of FREQUENCY

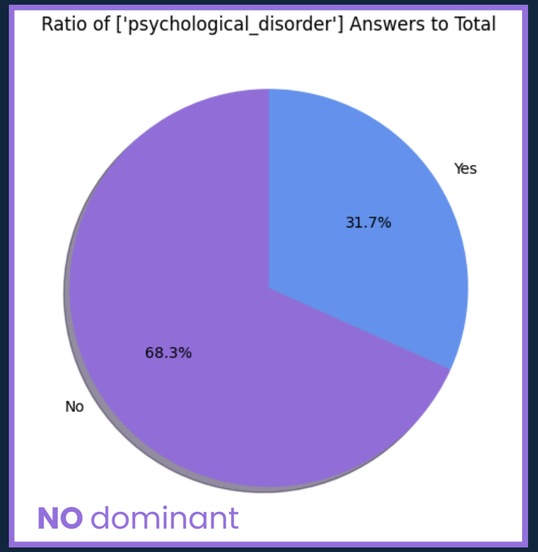
Revisiting these measures as I've mentioned in a past slide, I have explored the frequency of several categorical columns I have chosen to explore as per the Mental Health & Emotional Stability project.

By calculating the frequency, dividing it by the total number of values in a data, and multiplying it by 100, we get the ***relative frequency*** of the data. I've also used the technique of **pie charts** using the matplotlib tool to easily visualize the differences between the 'yes' and 'no' results.

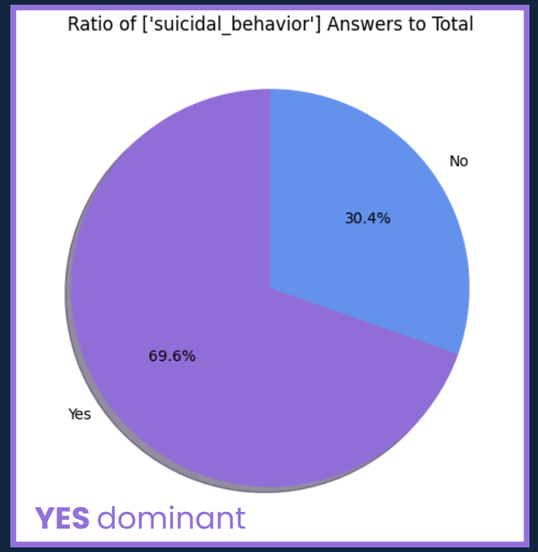
The data regarding irritable mood, aggressive behavior, & difficulty controlling emotions came to be a close 50/50, therefore I will not be including them. However, data regarding psychological disorders, hallucinations, unjustified anxiety/fear, drug or not, & suicidal behavior have clear differences.



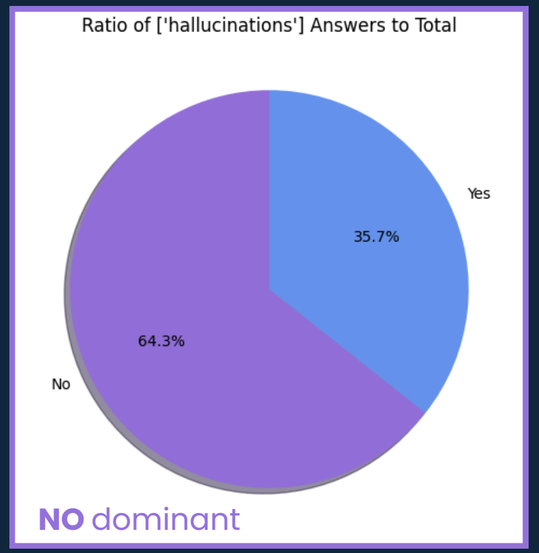
YES dominant



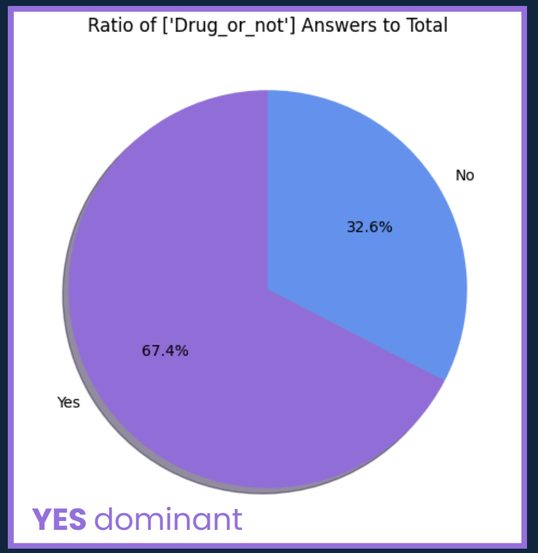
NO dominant



YES dominant



NO dominant



YES dominant

INSIGHTS

As it was plotted, the data of unjustified anxiety/fear, suicidal behavior, *and* drug or not lead with answers of ‘yes’ with over 60%. On the other hand, psychological disorders and hallucinations lead with answers of ‘no’ with over 60%.

DECISIONS

Therapy & Workshops for Those with Anxiety & Suicidal Tendencies -
Since it was presented that these 3 factors are the ones leading in people answering them as ‘yes’, it is incredibly important we host therapy sessions, workshops, and discussion groups for individuals with anxiety and suicidal tendencies.

Rehabilitation Therapy for Those who Take Drugs -
Those who have tested positive for drug abuse can attend a specialized rehabilitation process in order to help and aid them in becoming clean and rebuild their lives and health with professional psychiatrists and doctors.

How Did **Descriptive Analytics** Help?

By using the techniques of frequency, relative frequency, and pie charts, I was able to determine and visualize the larger chunk of answers from all the columns I have chosen to explore. Measures of frequency aid in understanding a little deeper the nature of the data in order to see what data or values are highly repeated and appearing the most. In this case, it allowed me to see the factors we *need* to prioritize to help those struggling with them.

Relationships Between Columns/Data

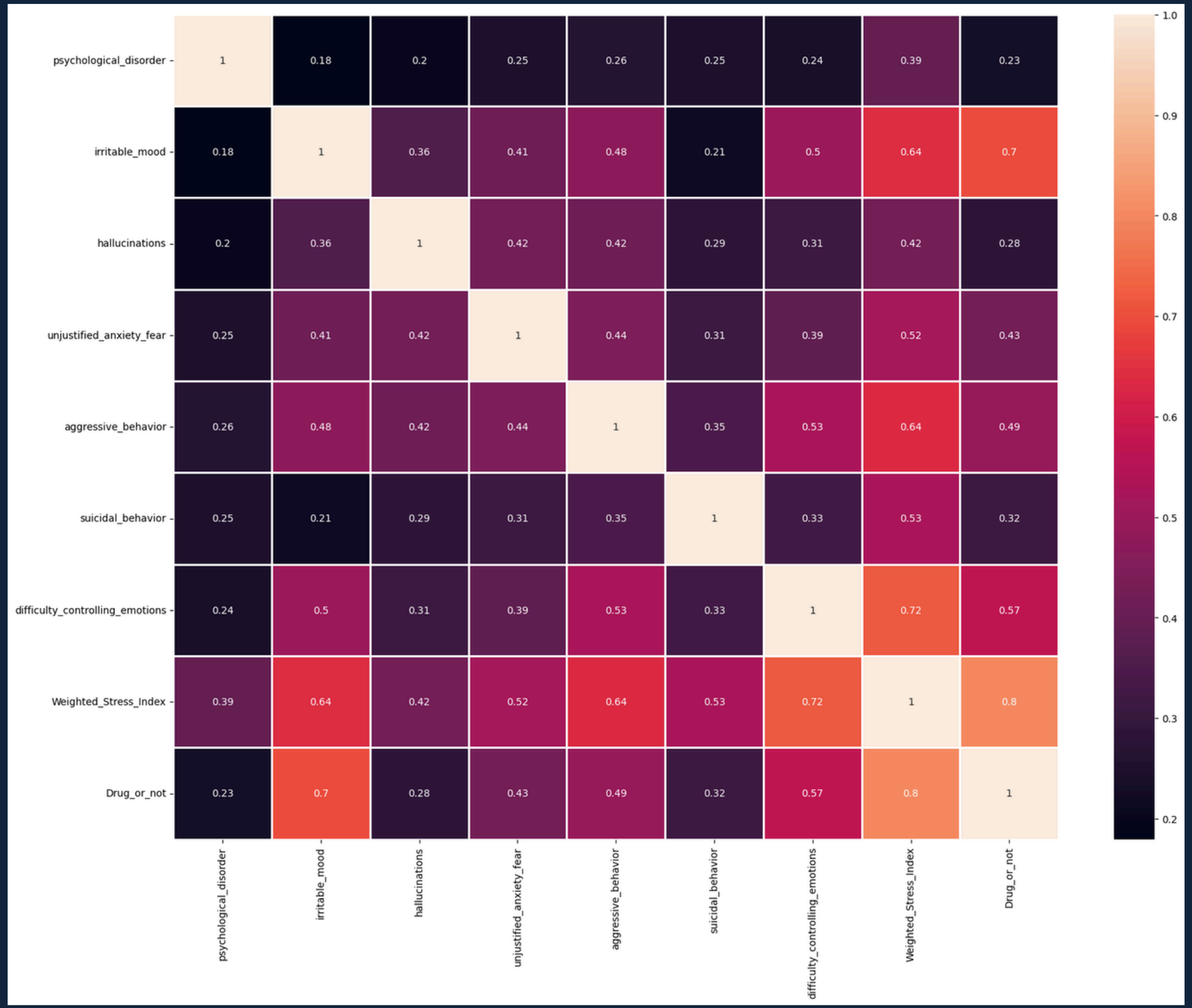
Correlation Matrix: Heatmap

How does a heatmap (correlation matrix) work and help in finding relationships? The heatmap calculates the **linear relationship** between two variables through the *correlation coefficient* equation. It represents said relation as levels from -1 to 1, with -1 being a strong, *inverse* relation and 1 being a strong, *direct* relation - if it's 0 then there is **no** relation. This process of finding relations between 2 columns helps find patterns and trends, so I've taken a threshold of 0.2 *and up* to focus on.

To achieve this technique, I've used 3 tools: sklearn label encoder to change categorical values to numerical, and seaborn & matplotlib to plot the actual heatmap. Another technique, .corr(), was also used to calculate the actual correlation of every 2 columns.

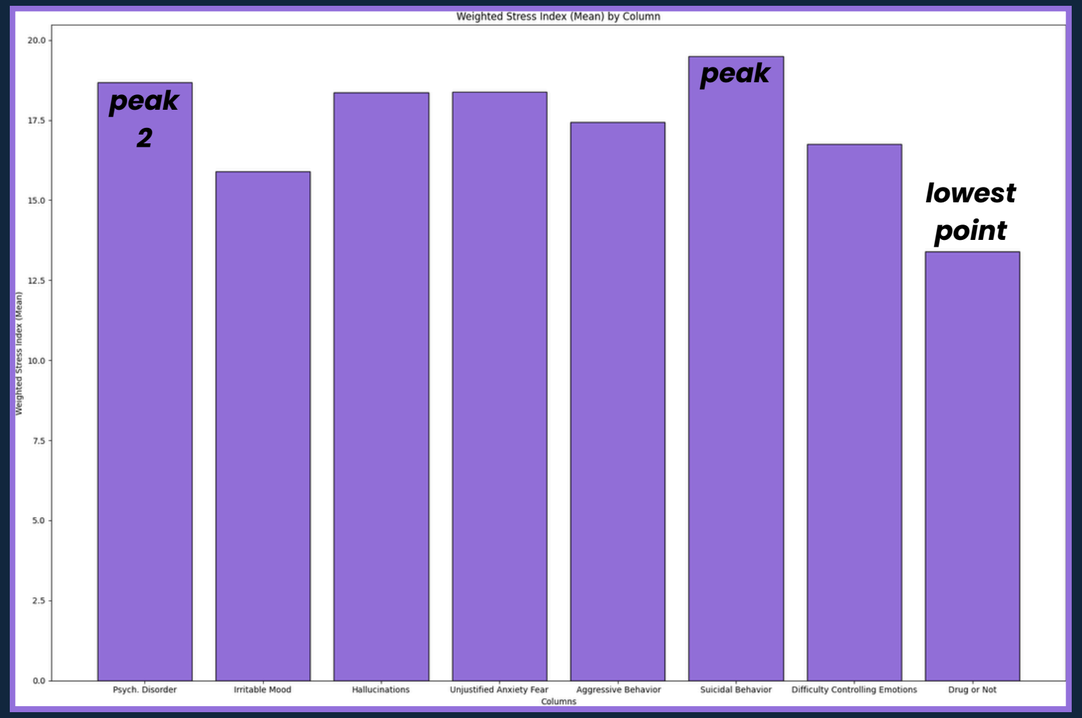
NOTABLE Relations To Explore Deeply:

1. irritable_mood & Drug_or_not = **0.7**
2. irritable_mood & Weighted_Stress_Index = **0.64**
3. aggressive_behavior & Weighted_Stress_Index = **0.64**
4. difficulty_controlling_emotions & Weighted_Stress_Index = **0.72**
5. Weighted_Stress_Index & Drug_or_not = **0.8**



GROUPBY: Weighted Stress Index

Note - *Techniques used are groupby and bar chart from matplotlib tool*




```
psychological_disorder
No     13.736834
Yes    23.612941
Name: Weighted_Stress_Index, dtype: float64
```

```
irritable_mood
No      8.365643
Yes    23.446425
Name: Weighted_Stress_Index, dtype: float64
```

```
hallucinations
No     13.165944
Yes    23.544606
Name: Weighted_Stress_Index, dtype: float64
```

```
unjustified_anxiety_fear
No     12.145748
Yes    24.613844
Name: Weighted_Stress_Index, dtype: float64
```

```
difficulty_controlling_emotions
No      8.357072
Yes    25.159556
Name: Weighted_Stress_Index, dtype: float64
```

```
aggressive_behavior
No      9.956120
Yes    24.901851
Name: Weighted_Stress_Index, dtype: float64
```

```
Drug_or_not
No      3.421351
Yes    23.373601
Name: Weighted_Stress_Index, dtype: float64
```

```
suicidal_behavior
No     12.799439
Yes    26.188835
Name: Weighted_Stress_Index, dtype: float64
```

Insights

It is important explore what factors causes the stress levels to increase the most, therefore I've used the groupby technique to find how each possible factor affects the mean of the stress.

Although the mean values are quite similar, there is still a significant difference. As we can see, suicidal behavior leads *however* in this case it mean it contributes the least. When suicidal behavior is equal to no, the stress index is *still high* meaning other factors are contributing greatly to the value. Following this logic, drug or not *affects the stress levels* the greatest. When it's 'no'„ stress levels are very low at ~3. but shoot up to ~23 when 'yes'. The data of difficulty controlling emotions, irritable mood, *and* aggressive behavior acts a similar way - when 'no', stress is relatively low but shoots up once 'yes'. So, hallucinations, psych. disorders, *and* anxiety contribute much less.

DRUG PREVENTION & REHABILITATION FOUNDATION

DECISIONS

Note: *We will prioritize those struggling with irritable mood, difficulty controlling emotions, aggressive behavior, and drug usage*

Therapy Sessions -
It is important to find the *root* cause of these present factors in such individuals and guide them to lead themselves a peaceful life.

Wellness App -
As a recognized foundation, we can create a platform that includes studied & proven ways to decrease strong, negative emotions. These ways can include calm sounds, white or brown noise, breathing exercises, brain exercising puzzles, etc.

Stress Management Skills -
It's important to note that certain behaviors are caused by *inheritance* or underlying physical disorders that may be difficult to immediately solve. In the meantime, workshops that teach stress management skills to reduce their stress levels.

Outdoor Activities -
It is scientifically proven that moving our bodies and truly experiencing the outside world can ease our moods - therefore, holding camp and activities can ease people's burdens.

How Did **Descriptive Analytics** Help?

By using the techniques of groupby, mean, *and* bar chart (with the help of the matplotlib tool to plot it), I was able to quickly identify the factors that contributed to the stress levels the most.

The groupby function easily divides the categorical columns into their values of 'yes' and 'no' with the corresponding mean of the weighted stress index. This allowed me to easily learn how the average stress value is affected when a certain factor is 'yes' or 'no'. *If* the stress value was already high when the factor is at 'no', this indicates quite the weak relation as other factors are affecting the stress. *However*, if the stress value was low when the factor was at 'no' and shot up once it was 'yes', this means it's a *heavy* contribution to the stress level.

Additionally, the groupby technique simple reiterated what the heatmap has shown - the strongest relations with the weighted stress index were as follows: *irritable mood* 0.64, *aggressive behavior* 0.64, *difficulty controlling emotions* 0.72, *drug or not* 0.8

Thus, creating decisions that *aim* at the people struggling with these factors is a priority to decrease their stress in one way or another as they suffer the highest amount of stress compared to other group. It's important to allocate resources well and work on decreasing the *highest* stress levels first and foremost.

DRUG PREVENTION & REHABILITATION FOUNDATION

GROUPBY & CROSSTAB: Drug or Not

It is important explore what factors contribute the greatest to whether or not a person is a drug user or not, therefore I've once again used the groupby technique *and* included crosstab to find how each possible factor affects outcome of drug or not.

irritable_mood	No	Yes	All
Drug_or_not			
No	69	5	74
	~93% no		
Yes	30	123	153
		~80% yes	
All	99	128	227

difficulty_controlling_emotions	No	Yes	All
Drug_or_not			
No	67	7	74
	~90% no		
Yes	45	108	153
		~70% yes	
All	112	115	227

Note - *The crosstab, or contingency table, simply shows the frequency of a combination of values between 2 categorical columns.*

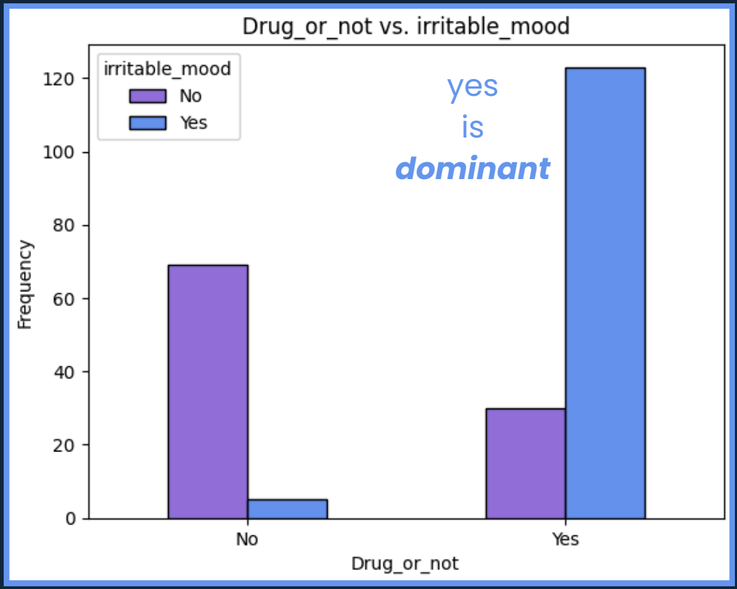
Insights

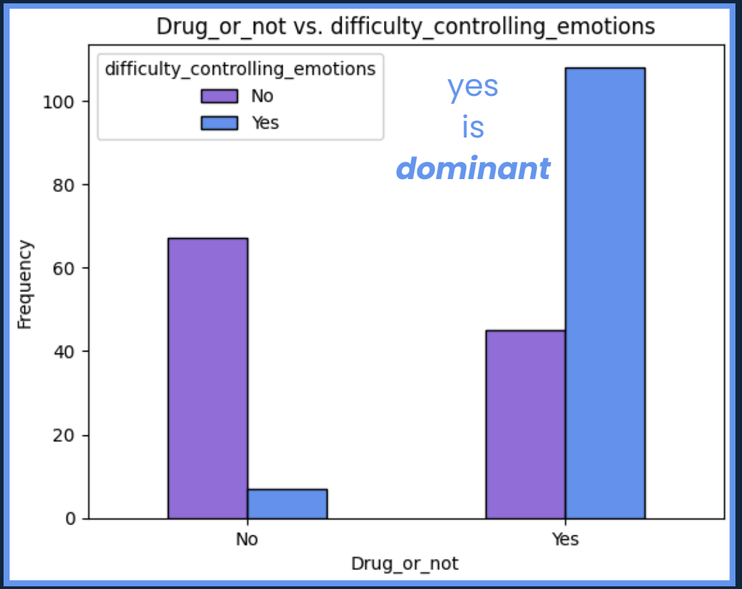
As it can be extracted from the groupby *and* crosstab results from the figures between drug or not with irritable mood *and* difficulty controlling emotions, these 2 columns *greatly* affect whether a person is a drug user or not. When *any* of these columns is 'no', the number of people taking drugs plummets a good amount. When they're both 'yes', the count of drug users increases. *However*, it is clear that irritable mood has a larger area of effect than difficulty controlling emotions from the percentages derived from the crosstab when they're = to yes (80% vs. 70%). This is also shown in the groupby figure on the right - when irritable mood is 'yes' but emotions is 'no', the frequency of drug takers is still quite high at 32. Whereas when the answers are reversed, the answer is *only* 14, which is still quite large but not as big.

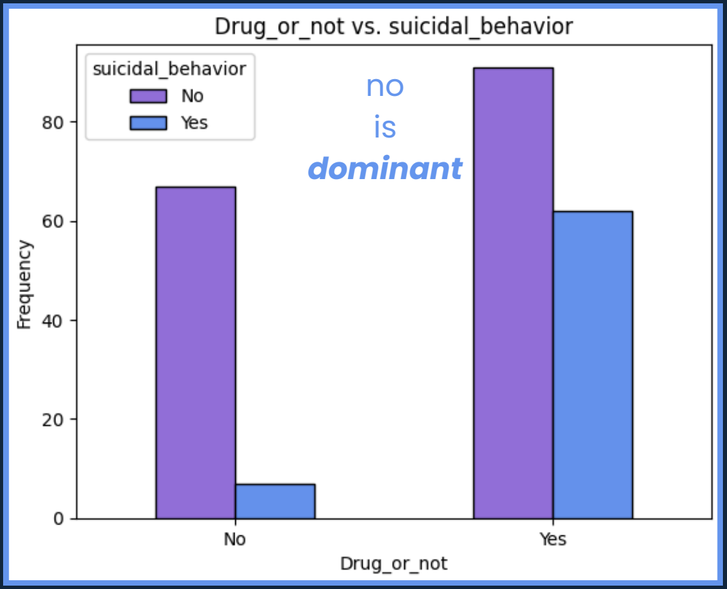
irritable_mood	difficulty_controlling_emotions	Drug_or_not	
No	No	No	63
		Yes	14
	Yes	No	6
		Yes	16
Yes	No	No	4
		Yes	31
	Yes	No	1
		Yes	92

It is important to note that **I did not include *all* the columns** in the dashboard as *only* irritable mood and difficulty controlling emotions produced clear cut results that indicate they *greatly* contribute to whether a person is a drug user or not.

CROSSTAB GRAPH:
Drug or Not







INSIGHTS

For visualization purposes, I've also used the technique of the bar chart (matplotlib as tool to plot the crosstab) in order to *really* see how much the frequency of 'yes' to drug usage is when irritable mood and difficulty controlling emotions is answered as 'yes'. For comparison purposes, I've also attached the bar chart of suicidal behavior to see the *difference*.

It is clear as day that the frequency of whether a person is a drug user or not is greatly affected by whether or not their mood is irritable, or they have a difficulty controlling their emotions. However, being a drug user *does not* necessarily lead to suicidal behavior as it's seen on the far right figure that 'no' in suicidal behavior is dominating in either case.

DRUG PREVENTION & REHABILITATION FOUNDATION

DECISIONS

Note: *This is aimed at people struggling with irritable mood and difficulty controlling emotions, prioritizing drug takers.*

Therapy & Counseling Sessions -
It can be incredibly difficult and may take a long time to become clean from drugs - but that does not mean they should not work on their emotions in the meantime. Therefore, as a drug focused foundation, we must open up specialized access for them to get therapy & counseling in order to reduce these negative emotions as best they can while in rehab. Non-drug takers may also benefit from this for a peace of mine and avoidance of potential drug abuse in the future.

Support & Discussion Group -
Feeling seen and being understood from people with a similar experience is incredibly important, especially for the drug takers. A support & discussion group is encouraged to be formed to share stories, emotions, and feelings in a non-judgmental environment to release negativity and learn new things.

Stress Mangement -
Aimed mainly at non-drug takers, it's important for them to learn to obtain healthy coping mechanisms and manage their mood in a peaceful manner. This is crucial as there *may* be a high chance of future drug usage as a way to cope - we must look after them and guide them away from this path.

How Did **Descriptive Analytics** Help?

By using the techniques of groupby, crosstab, *and* count, I was able to identify clear patterns & trends to find what affects the usage of drugs the most.

The groupby function easily divides the categorical columns into their values of 'yes' and 'no' with the corresponding count of the drug or not data. As mentioned in the previous slide, I was able to find that when irritable mood and difficulty control emotions data was recorded as 'no', drug usage was low. But, when they were recorded as 'yes', drug usage was high. The crosstab technique also reiterated what the groupby had shown, further proving that the percentage of 'yes' drug takers was much higher than 'no' when either of the 2 columns were also 'no'. The bar chart technique graphing the crosstab relation allowed me to truly visualize the dramatic effect of those 2 columns.

Once again, these strong relations were indeed spotted in the heatmap earlier, showing drug or not with irritable mood at 0.7 and with difficulty controlling emotions at 0.57

So, since these 2 are the *biggest* factors towards drug usage, it's important to come up with decisions & solutions aimed at them. Both drug takers and non-takers benefit from the suggested solutions to better their moods *and* help in avoiding drug usage to cope.

References.

Kalyan. (2017). *5 Core Activities of Data Analysis | Epicycles of Data Analysis*. MAKE ME ANALYST. Available at: https://makemeanalyst.com/5-core-activities-data-analysis-epicycles-data-analysis/#:~:text=There%20are%205%20core%20activities%20of%20data%20analysis%3A. (Accessed: 17 April, 2024).

Hiller, W. (2022). *A Step-by-Step Guide to the Data Analysis Process*. Career Foundry. Available at: https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/. (Accessed: 17 April, 2024).

Arora, S.K. (2020). *What is Data Analysis? Methods, Techniques & Tools.* Hackr.io. Available at: https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools. (Accessed: 17 April, 2024).

Glenn, N. (2020). *Data Analysis: Types, Methods & Techniques (a Complete List)*. Analyst Answers. Available at: https://analystanswers.com/data-analysis-types-methods-techniques-intro-for-researchers-corporates/. (Accessed: 17 April, 2024).

Trevelyan, L. (2023). *UNODC report shows significant increase in drug use as international responses diverge.* Available at: https://www.ibanet.org/unodc-report-drug-use-increase. (Accessed: 19 April, 2024).

Stein, S. (2019). *Seasonal Affective Disorder and Substance Abuse*. Available at: https://americanaddictioncenters.org/treating-depression-substance-abuse/sad. (Accessed: 20 April, 2024).

QuestionPro Collaborators. (2023). *Correlation Matrix: What is it, How It Works with Examples*. QuestionPro. Available at: https://www.questionpro.com/blog/correlation-matrix/. (Accessed: 17 April, 2024).

Zach. (2021). *What is Considered a Good Standard Deviation?.* Statology. Available at: https://www.statology.org/what-is-a-good-standard-deviation/. (Accessed: 18 April, 2024).

Lawton, G. (2020). *8 Types of Bias in Data Analysis and How to Avoid Them*. SearchBusinessAnalytics. Available at: https://www.techtarget.com/searchbusinessanalytics/feature/8-types-of-bias-in-data-analysis-and-how-to-avoid-them. (Accessed: 20 April, 2024).