-

TDW6323/TDWF6323 Data Wrangling and Visualisation

Trimester 3 2024/2025 (Term 2420)

Faculty of Information Science & Technology

PROJECT REPORT

GROUP NAME

NO	NAME	STUDENT ID	SIGNATURE
1	ANIS SYIFAA' BINTI MOHD ZAFFARIN	1211112369	
2	KUEH PANG TENG	1211112304	
3	NUR INSYIRAH IMAN BINTI MOHD AZMAN	1211112312	
4	SOFIA BATRISYIA BINTI MOHAMAD FARIS	1211111880	

Table of Contents

Dataset Description

The Sleep Health and Lifestyle Dataset consists of 400 rows and 13 columns, namely ID, Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, and Presence of Sleep Disorders. This dataset aims to study sleep quality and its relationship with one's lifestyle. This dataset provides insightful information that is related to the objective of the study.

Each subject has a unique ID associated to identify any repetitive rows. Gender, age and occupation data is collected to see if any of these biological factors influence their sleeping patterns. As for sleep health, we collected the subjects' sleep duration and their sleep quality (ranging from 0 to 10). The lifestyle factors are examined through physical activity levels (ranging from 0 to 100), stress levels (ranging from 0 to 10), and their BMI categories. We also analyze their cardiovascular health by collecting their blood pressure, heart rate measurements and daily steps. Last but not least, we recorded the presence of any sleep disorder in each subject. 'None' in this column indicates the person does not have any sleep disorder. 'Insomnia' indicates the person is having trouble sleeping at night or having trouble staying asleep, thus affecting the sleep quality. 'Sleep Apnea' means that the person is having trouble breathing while sleeping, which could affect his or her sleep quality and indicates a potential health risk.

We chose this dataset to study the relationship between lifestyle factors such as gender, occupation, age, physical activity, sleep duration and daily steps affect one's quality sleep. We are also taking physical and mental health such as stress level, BMI categories, blood pressure, and heart rate into account. Hence, we proposed 4 interesting questions to deepen our understanding regarding the collected data. Here is the link for the dataset used in this project: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

1. Does BMI influence ones' sleep disorder?
2. How do BMI categories correlate with blood pressure and heart rate?
3. Are there any trends or relationships between sleep efficiency and stress level?
4. Which occupations experience the highest and lowest quality of sleep?

Literature Study of Existing EDA Approaches Across Different Industries

HEALTHCARE

In a research paper titled Global Rates of Glaucoma Surgery mainly focuses on studying the number of glaucoma surgery done all over the world. The author provided an overview of the number of surgeries in different regions and countries. In the paper, the author mentioned that the significant differences in the number of glaucoma surgery rate (GSR) itself is challenging, as they need to further study the cause of the differences (Kaweh Mansouri, 2013). He used scatter plot to represent his study. He broke the data collected into a few categories to deepen the study. This study, however, has a few disadvantages. First, it is a lack of data as only 38 countries were able to provide the information. Thus, this study might not be 100% accurate. Second, the reliability of the data is also a concern as some data were collected from leading ophthalmologists, rather than the national ophthalmologists' societies, which could be subject to bias. Last but not least, the study was not done in a population of glaucoma people, which could lead to inaccuracy of inference.

The study of student stress brought on by schools along with external circumstances is presented in the article 50 Current Student Stress Statistics: 2024 Data, Analysis & Predictions. This study describes the factors that influence students' stress levels in K–12 schools, colleges, and universities. The study also includes a pie chart illustrating the stress of students in colleges and universities. The analysis shows that 45.1% of students experience stress levels above normal and 9% of students experience no stress or stress levels below average (Bouchrika, 2024). There is also a bar chart of the common factors pressuring teens including getting good grades, physical appearance, and fitting in society. However, the data verified that students typically manage their stress on a personal level instead of consulting a counselor. The EDA is easily understood by a wide audience and clearly shows the distribution of students' stress levels, while the bar chart highlights the factors causing the stress. However, the visualization can also cause bias and misleading as the authors might overlook other stressors and the label is not well-differentiated, especially in the bar chart potentially leading to an incorrect conclusion.

The Births, Economic Growth, Mortality and Murder in a Developing Country article shows the study of the birth rate and the factors of inflation, GDP per capita growth, and mortality (Bourne, 2011). It also shows the correlation between birth and negative factors such as murder, unemployment, and annual exchange rate. The EDA used are scatter plots that show the average, annual, and log registered births over different periods, effectively showing an insight that influences the birth rates and how economic and societal issues can impact fertility. For example, the analysis shows that periods of economic hardship are associated with higher birth rates, while exchange rates and murder have a negative impact. However, the relationship discovered in this study is complex to interpret and might overlook other important factors such as education and healthcare.

The article Health Problems of Victims Before and After Disaster: A Longitudinal Study in General Practice studies the long-term health issues that may have arisen following the explosion in the Netherlands. The study involved 9329 victims and a control group of 7392 tracked over time by using electronic medical records (C Joris Yzermans, 2005). The control group was compared with victims for the pre- and post-disaster. The graphs show three different lines relocated lines, non-relocated victims, and the control group presenting psychological problems and medically unexplained physical symptoms (MUPS). The large sample size can effectively find and generalize the statistics and provide trustable insights into the impact on health. On the downside, depending on general practitioners' records, all health issues may not be recorded, and the media attention that they put on these issues might influence the results during specific periods.

The study "Mental Disorder Classification With Exploratory Data Analysis (EDA)" analyses mental health data from multiple perspectives and sources to determine the relationships between mental disorders and other variables such as characteristics, prevalence, and distribution. One major challenge is the classification of mental disorders based on symptoms as the classification may not correspond to each other, and can differ across cultures, contexts and time periods. EDA is used to compare different mental disorders classifications methods, focusing on 17 critical symptoms for diagnosing Bipolar Disorder and Major Depressive Disorder and the individuals without these conditions are distinguished. The results are summarised in histograms including both the categorical approach which view mental disorders as distinct entities, and the dimensional approach which view them as a spectrum. It is also concluded that EDA allows several practical insights in the medicine by helping doctors to better understand their patient's conditions, monitoring them in real time and deliver timely treatment. It also assists policymakers by offering data-driven insights for efficient resource allocation and targeting areas according to the highest need for mental health services. EDA enhances the accuracy and depth of mental health assessments in a non-invasive method, making it suitable for monitoring the patients regularly without causing discomfort. (Juanto Simangunsong, 2024)

The article titled “Exploratory Data Analysis (EDA) methods for healthcare classification” aims to understand the population served by a healthcare facility and determine the prevalence of various diseases or conditions within a patient population through EDA. The billing amount based on the medical conditions and medications are clearly presented through multiple bar charts. This study analyses a dataset of 10000 observations, and providing insights into patient characteristics, with ages ranging from 18 to 85 and an average of 51. Healthcare costs range from 1000.18 and 49995.90 with an average bill of 25516.80 and a standard deviation of 14067.29. It also shows that 25% of patient had bills below 13506.52 and 75% below 37733.91, while inpatient room numbers range from 101 to 500. This data overview provides a foundation for healthcare analysis and decision making. They are useful in understanding the patient demographics, disease prevalence, resource allocation. It also reveals the trends and patterns of disease rates or healthcare utilisation over time which is important to support the development of advanced predictive models in increasing the capacity of healthcare systems to provide a more responsive and personalised care. (Hanna Willa Dhany, 2023)

The article entitled “Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020” analyses the trends and patterns of breast cancer incidence and mortality worldwide. The study provides a comprehensive analysis of breast cancer statistics across various countries and regions by utilizing the data from multiple cancer registries. Hence, a significant difference in breast cancer incidence and death rates is found between developed and developing countries, which developed countries had higher incidence rates, due to factors such as lifestyle, reproductive habits, and screening accessibility. However, there is also some of problems faced in this study, such as the use of data from various resources may cause issues with data comparability. Differences in data collection techniques and definitions of cancer cases could skew the results. Therefore, in this case, exploratory data analysis, or EDA, is used to assist researchers in understanding data patterns, so that they could interpret the results accurately and draw informed conclusions. (Shaoyuan Lei, et al., 2021)

The article entitled “A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050” presents a meta-analysis of global food projections, which focuses on the risk of hunger. The study utilizes the findings from various research studies to provide a thorough overview of food security estimates and assess the risk of hunger across the globe. A systematic literature review is conducted by extracting data from 57 studies and categorised based on their methodologies. Next, the analysis utilized graphical techniques and meta-regression model to analyse the extracted data, then identify trends over time, correlations between variables, and outliers. As a result, the results obtained indicate a strong overlap between the analysed studies and the estimated confidence bands from the meta-regression, ends up in suggesting minimal selection bias. (Michiel van Dijk, Tom Morley, Marie Luise Rau, & Yashar Saghai, 2021)

In the article "Heart Disease Prediction using Exploratory Data Analysis", the authors discuss about the application of Exploratory Data Analysis (EDA) in predicting heart disease. The study uses the K-means clustering algorithm to analyse a dataset that contains 209 records with 8 attributes, which include age, chest pain type, blood pressure, and heart rate. As a result, the predictions made using EDA and the K-means algorithms are accurate, which clearly prove the effectiveness of these methods in the healthcare industry. Different kinds of graphical representations are also used in this study such as histogram and pie charts to provide a visual representation of the data, which helps in understanding the finding results better. (R. Indrakumari, T. Poongodi, & Soumya Ranjan Jena, 2020)

The article titled "Energy Consumption Optimization of a Fluid Bed Dryer in Pharmaceutical Manufacturing Using EDA (Exploratory Data Analysis)" discuss about the implementation of using EDA in the optimization of energy consumption of fluid bed dryers. The main goal of this study is to reduce the preheating time and energy consumption of the fluid bed dryers. Hence, EDA is used to analyse over 700,000 data records from 56 sensor signals caught at one-minute intervals, which helps to identify trends and systemic behaviours within the fluid bed drying process. As a result, the results concluded that it is possible to reduce the machine's preheating time by more than 50% to achieve an average reduction of one hour per batch. This proves the effectiveness of EDA in pharmaceutical manufacturing processes, which shows its potential for other equipment and processes within the industry. (Roberto Barriga, Miquel Romero, Houcine Hassan, & David F. Nettleton, 2023)

SPORTS

The article A Medal in the Olympics Runs in the Family: A Cohort Study of Performance Heritability in the Games History main purpose is to study whether having a genetical history of a former Olympic champion in the family will increase the chance for an individual to become a champion too. This study includes 125,051 worldwide athletes that has participated in Olympics between 1896 to 2012 (Juliana Antero, 2018). The data was collected from a reliable database. They group the subjects into two categories: kinship with former non-medallists and kinship with former medallists. The chi-square test was used to gain the statistical analysis. They visualised the analysis in the form of a bar graph for clearer interpretation. A line graph was used to summarize the relationship of frequencies of medallists with the gap of participation years between them and their kinship. By using the line graph, they can make a conclusion easier. However, the author thinks that his method in this research were lacking subjects and that could lead to hasty generalization to all populations of athletes. All in all, this paper could be used as a springboard to conduct further research.

The sports industry also influences the development of the economy will be explored in this paper entitled The Analysis and Research on the Influence of Sports Industry Development on Economic Development. The first EDA used is an S-curve graph that shows the research literature growth over time, displaying steady increases in the sports industry from 2013 to 2020 (Kim, 2022). The study also includes a regression analysis using a regression correlation plot to determine the factors influencing economic development in the sports industry including economic growth, types, economic assessment, action time, and scale. The big data used in this study deeply influenced the result found and showed the growth of different phrases in the sports industry meanwhile the regression plot pinpointed the most factors such as scale and action time. Thus, it simplifies the data by visualizing using the EDA and making it more accessible to a wide audience. On the other hand, the linear relationship between variables needs to be assumed linear, which might simplify the potential non-linear relationship. The growth assumption can also be mistaken if some external force or phenomenon can disrupt the entire conclusion.

Exploratory Data Analysis (EDA): A Study of Olympic Medallist is an article that focus to analyse the changes of data of Olympic Medallist from 1896 to 2014 through univariate, bivariate, and multivariate analysis. It provides detailed information and explore the relationships between the variables in the data. The data analysed includes the aspects of the Olympics such as the country-specific data like the country name and code, total population and GDP per capita and the data of the Olympics such as the year, city, sports, disciplines, the athlete’s name and the medals won, including both the summer Olympics and winter Olympics. Python is used for the EDA with Visual Studio Code as the platform in this study. In the univariate analysis, the gender variable is selected, revealing that men athletes have more historical accomplishments than female athletes by 268%. The Bivariate analysis used the Chi-Square Test of Independence to determine the relationship between the medals and year and it was concluded that both the variables are not independent. For the multivariate analysis, the Pearson’s Correlation is used to determine the relationship between the country, medal achievements, medallist by gender, medals count, the country and sport type. It shows that there is a positive correlation between bronze, gold and silver medals and the types of sports in which they were won. In conclusion, using the EDA approach to analyse the data provides a deeper understanding of the factors influencing the Olympics athletes across the country. (Noviyanti T M Sagala, 2022)

EDUCATION

In the paper titled The Compatibility Student Choice of University Majoring: A Preliminary Studies, the authors aimed to study the suitability of Universitas Negeri Padang, Indonesia students majoring choices with their interest. It included a total of 122 students and all of them were randomly distributed. They visualized the percentage of the students' suitability of the major they are occupying in a bar graph. 65% of the students felt they are taking the current major of their interest. In the survey, they also asked the students whether they have preliminary information regarding the majors they are going to take in university. Surprisingly, almost 90% of them already have some insights of the chosen major. Given that 32% of the students are not really satisfied with their major in university, there must be other reasons influencing the result of the study. Therefore, more information needs to be gained so we can have a precise conclusion. (Daharnis, 2016)

Another paper titled Cause Analysis of Students Dropout Rate in Higher Education was selected to help study the application of EDA in the education industry. This paper aims to study the reasons of dropout cases of first year students in Latvia University of Agriculture. The study focuses on engineering students. They analyzed a few factors that might be influencing the dropout cases such as gender, priority chosen programs, secondary school grades, and finance. A little bit of EDA is applied in the study. A pie chart is included in the article to show the percentage of students who continue their study and those who drop out of university. 64.1% of them continue their study into the second year. A total of 34.4% have left university in the first year, and only 1.5% of the sample have never initiated their study. The bar chart in the article showed the mean secondary schools of the dropout students and those who continue their study to second year. They reached a conclusion that those who gained lower marks in secondary schools are more likely to drop out of university. (Liga Paura)

The article titled “Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis”, aims to determine the student performance patterns in engineering and identify any correlations between various variables in the dataset and student performance in engineering. This study highlighted a gender disparity in engineering enrolment and the results revealed that although women have lower enrolment rates, but they outperform men academically. A notable issue is the drop out of some students due to choosing the wrong qualifications, underscoring the need for better career guidance and a recommender system to assist students in choosing the appropriate qualifications. The article also urges universities to reconsider the completion of experiential learning requirements impacted by COVID-19. They also show that mainstream program students perform slightly better than those in extended programs due to stigma around extended programs. The findings suggest intervention using student performances patterns for tailored support to address for South Africa’s low graduation rates. The results are presented in bar chart, correlogram, pie chart, heatmap, box plots for better visualisation. In short, the results of the study aligned with the findings from the EDA approach. It increases the efficiency of study by presenting clear information, making it easy to understand which simplifies the analysis of the findings. (MFOWABO MAPHOSA, 2023)

INDUSTRY: HEALTHCARE

TITLE	ADVANTAGES	DISADVANTAGES
Global Rates of Glaucoma Surgery (Kaweh Mansouri, 2013)	- Able to deepen the study based on the data collected. Leads a clearer path for new research opportunity	- Unreliable data source - Lack of resources
Mental disorder classification with exploratory data analysis (EDA) (Juanto Simangunsong, 2024)	- Reduce the subjectivity of self-reported symptoms through measurement of physiological responses - Allows real-time monitoring of emotional and stress responses, providing continuous data to track changes over time and in response to treatment - Minimise patient's discomfort during regular monitoring - Able to reveal hidden patterns within the datasets and offer new insights into factors and progression of the disorders	- High variability of data limiting ability to identify consistent patterns - Lack of diversity and small data sample may reduce the generalisation of findings - Missing values on key variables limit the ability to draw reliable conclusions
Exploratory Data Analysis (EDA) methods for healthcare classification (Hanna Willa Dhany, 2023)	- The information can help identify the prevalence of different disease or conditions within a patient population - Shows the process of patients in the healthcare system from diagnosis to treatment to follow-up to improve the patient care and reduce waiting time - Divide the patient population into different risk groups and allow personalised healthcare interventions - reveal patterns of adverse events related to treatments or medications,	- Requires further modelling techniques to offer predictive insights and forecast healthcare outcomes - EDA relies on visual representations that can be subjective that may require an expertise in the field to interpret the results accurately
Births, Economic Growth, Mortality and Murder in A Developing Country (Bourne, 2011)	- Valuable insights on the correlation	- The interpretation complex between the variables - May overlook other factors

Health Problems of Victims before and After Disaster: a longitudinal study in General Practice (C Joris Yzermans, 2005)	- More accurate results because of the large sample size	- Influence on the media attention - Dependency on GP records
50 Current Student Stress Statistics: 2024 Data, Analysis & Predictions (Bouchrika, 2024)	- The EDA can be understood easily - The distribution of data can be seen clearly	- Selective representation - Misleading visualization
Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020 (Shaoyuan Lei, et al., 2021)	- Identify underlying patterns, trends, and anomalies - Detect issues such as missing values or outliers	- Time consuming
A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050 (Michiel van Dijk, Tom Morley, Marie Luise Rau, & Yashar Saghai, 2021)	- Identify trends over time, correlations between variables, and outliers - Help researchers understand the data's structure, distribution, and potential anomalies	- Too much information in a graph can overwhelm the viewer
Heart Disease Prediction using Exploratory Data Analysis (R. Indrakumari, T. Poongodi, & Soumya Ranjan Jena, 2020)	- Help researchers to understand the underlying structure of the data - Provide visual representation of the data that is easy to interpret	- The using of large datasets can lead to time-consuming
Energy Consumption Optimization of a Fluid Bed Dryer in Pharmaceutical Manufacturing Using EDA (Exploratory Data Analysis) (Roberto Barriga, Miquel Romero, Houcine Hassan, & David F. Nettleton, 2023)	- Identify trends and systemic behaviours within the fluid bed drying process	- Time-consuming when dealing with extensive datasets

INDUSTRY : SPORTS

TITLE	ADVANTAGES	DISADVANTAGES
Exploratory Data Analysis (EDA): A Study of Olympic Medallist (Noviyanti T M Sagala, 2022)	- Provides better insights with different visualisation techniques and give insightful information - Understanding the correlation between attributes - Python is utilised for EDA because it is easy to learn, flexible, has a broad library and free	- EDA is subjective because the analyst's perspective can influence the analysis - Time-consuming for large datasets
A Medal in the Olympics Runs in the Family: A Cohort Study of Performance Heritability in the Games History (Juliana Antero, 2018)	- Easier to achieve a conclusion of the study - Reliable source without web scraping - Line graph makes easier to observe data	- Lack of subject of research that could lead to hasty generalization
The Analysis and Research on the Influence of Sports Industry Development on Economic Development (Kim, 2022)	- Easier identification of growth pattern - Pinpoint the analysis of the correlation between factors	- Over-simplification - Misinterpretation and fault conclusion

INDUSTRY : EDUCATION

TITLE	ADVANTAGES	DISADVANTAGES
Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis (MFOWABO MAPHOSA, 2023)	- Discover the patterns and correlations between student performance and other different variables - Visual representation of data makes it light and easy to understand about the current situation - Useful for analysing large datasets and gain valuable insights	- Results may be questioned for their accuracy, and might be extended to cover a broader varied student's sample - High cost for generating the results - More variables are needed to increase the reliability of the findings and understand the causal mechanisms between variables

	- Simplifies complex data, making it easier to understand	
The Compatibility Student Choice Of University Majoring; A Preliminary Studies (Daharnis, 2016)	- Sample study makes it easier to handle data - Simple and clear interpretation of data in the form of bar graph.	- Lack of information. - Should add more question in the survey for a more precise study.
Cause Analysis of Students' Dropout Rate in Higher Education Study Program (Liga Paura)	- Clear interpretation of data using pie chart and bar chart.	- Sample limitations - Focuses on short-period trends, might miss long-period observation that could contribute more to the study.

Data Wrangling

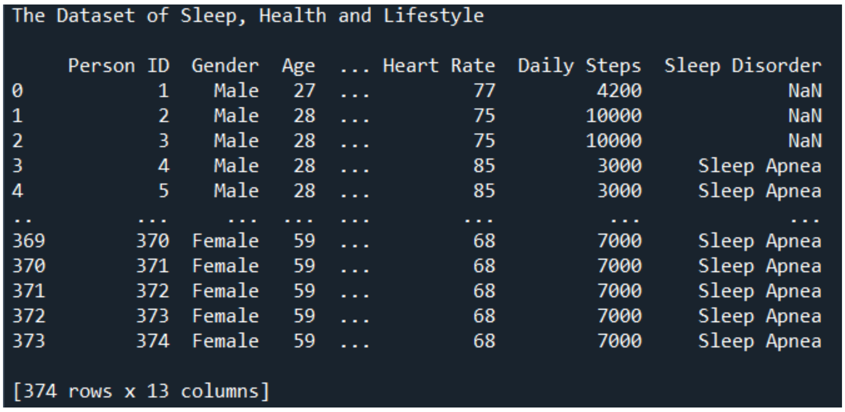
Using Pandas, load the dataset into a DataFrame using the command below by specifying the file path. This can create a structural table to easily access and manipulate the rows and columns for any further analysis.

import pandas as pd

df = pd.read_csv(""C:\\1211112312\\Bachelor\\2nd Year\\T2420\\TDW6323 DATA WRANGLING & VISUAL\\Project\\dataset\\sleep_health_and_lifestyle.csv"")

print(df)

The output:



```
The Dataset of Sleep, Health and Lifestyle

     Person ID  Gender  Age  ... Heart Rate  Daily Steps  Sleep Disorder
0            1    Male   27  ...         77         4200             NaN
1            2    Male   28  ...         75        10000             NaN
2            3    Male   28  ...         75        10000             NaN
3            4    Male   28  ...         85         3000     Sleep Apnea
4            5    Male   28  ...         85         3000     Sleep Apnea
..         ...     ...  ...  ...        ...          ...             ...
369        370  Female   59  ...         68         7000     Sleep Apnea
370        371  Female   59  ...         68         7000     Sleep Apnea
371        372  Female   59  ...         68         7000     Sleep Apnea
372        373  Female   59  ...         68         7000     Sleep Apnea
373        374  Female   59  ...         68         7000     Sleep Apnea

[374 rows x 13 columns]
```

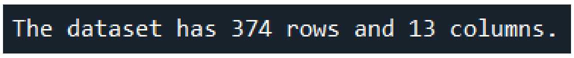
To see the total rows and columns, we can use the *df.shape* attribute as such:

rows, columns = df.shape

rows_columns = "\nThe dataset has %d rows and %d columns." % (rows, columns)

print(rows_columns)

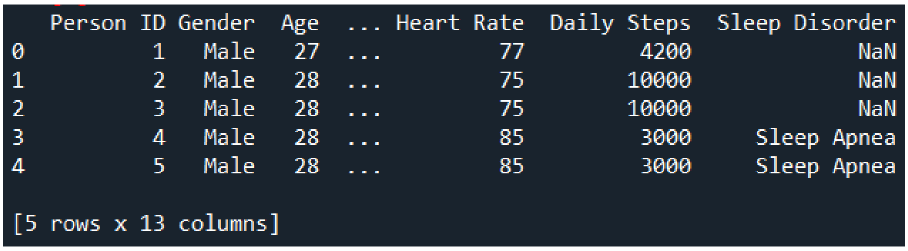
The ouput:



```
The dataset has 374 rows and 13 columns.
```

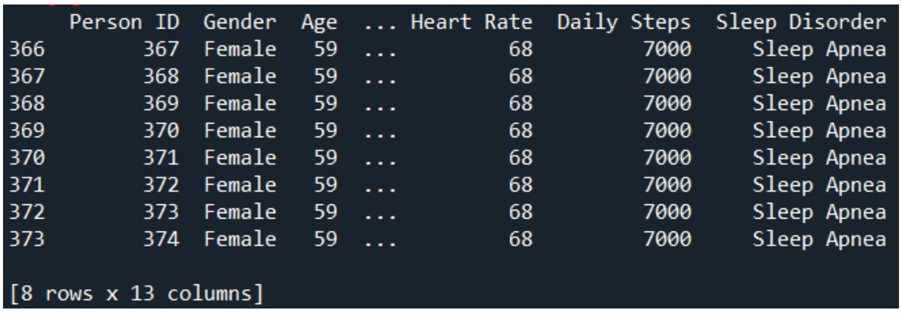
To see the list of rows by default, the *df.head()* function can display the first five, meanwhile, the *df.tail()* function can display the last five rows. To specify the number of rows to display, we can also add a number between the columns such as *df.tail(8).*

The output of the df.head() function:



```
   Person ID Gender  Age  ... Heart Rate  Daily Steps Sleep Disorder
0          1   Male   27  ...         77         4200            NaN
1          2   Male   28  ...         75        10000            NaN
2          3   Male   28  ...         75        10000            NaN
3          4   Male   28  ...         85         3000    Sleep Apnea
4          5   Male   28  ...         85         3000    Sleep Apnea

[5 rows x 13 columns]
```

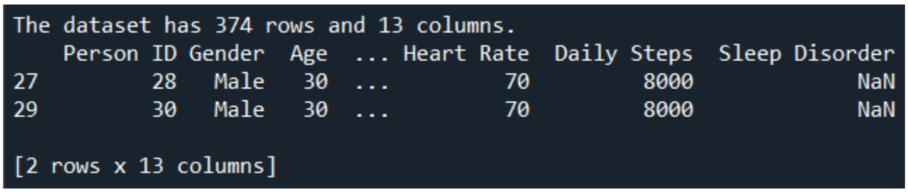
The output of the df.tail(8) function:



```
     Person ID  Gender  Age  ... Heart Rate  Daily Steps  Sleep Disorder
366        367  Female   59  ...         68         7000     Sleep Apnea
367        368  Female   59  ...         68         7000     Sleep Apnea
368        369  Female   59  ...         68         7000     Sleep Apnea
369        370  Female   59  ...         68         7000     Sleep Apnea
370        371  Female   59  ...         68         7000     Sleep Apnea
371        372  Female   59  ...         68         7000     Sleep Apnea
372        373  Female   59  ...         68         7000     Sleep Apnea
373        374  Female   59  ...         68         7000     Sleep Apnea

[8 rows x 13 columns]
```

To display specific multiple rows, *print(df.iloc[[27,29*]]) has been used to show rows at index 27 and 29.

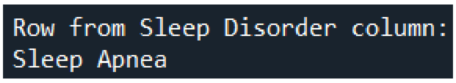
The output:



```
The dataset has 374 rows and 13 columns.
    Person ID Gender  Age  ... Heart Rate  Daily Steps  Sleep Disorder
27         28   Male   30  ...         70         8000             NaN
29         30   Male   30  ...         70         8000             NaN

[2 rows x 13 columns]
```

To display rows from specific columns using the *print(df['Sleep Disorder'][4])*, column Sleep Disorder row number 4.

The output:



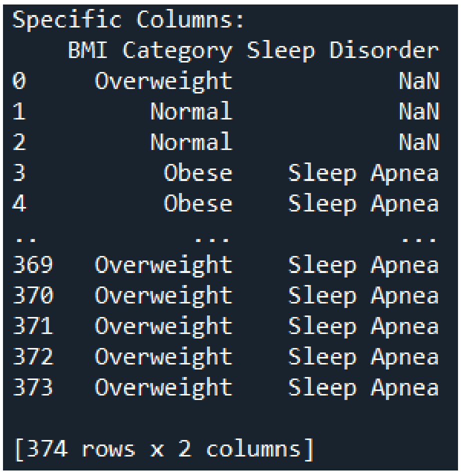
```
Row from Sleep Disorder column:
Sleep Apnea
```

For columns, we can also display multiple specific columns by using this command:

print("\nSpecific Columns:")

print(df[['BMI Category','Sleep Disorder']])

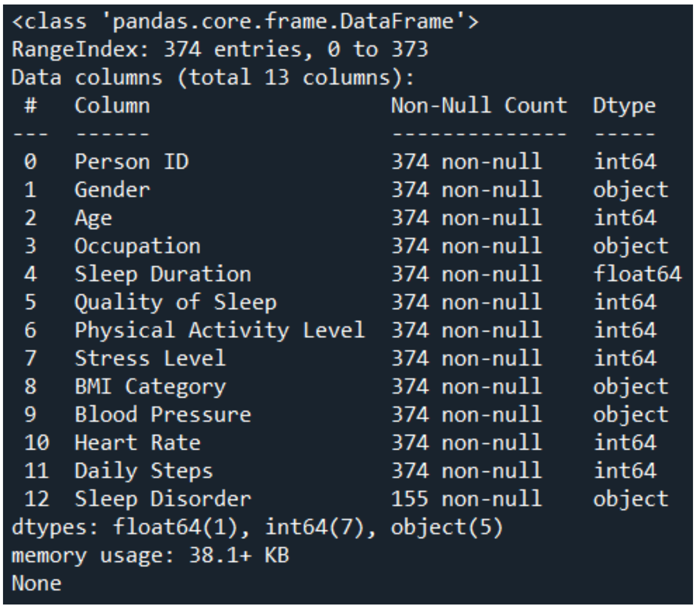
The output shows BMI Category and Sleep Disorder columns:



```
Specific Columns:
    BMI Category Sleep Disorder
0     Overweight            NaN
1         Normal            NaN
2         Normal            NaN
3          Obese    Sleep Apnea
4          Obese    Sleep Apnea
..           ...            ...
369   Overweight    Sleep Apnea
370   Overweight    Sleep Apnea
371   Overweight    Sleep Apnea
372   Overweight    Sleep Apnea
373   Overweight    Sleep Apnea

[374 rows x 2 columns]
```

For data types, the *df.info()* function will display the information of DataFrame and the data types of each column.

The output:



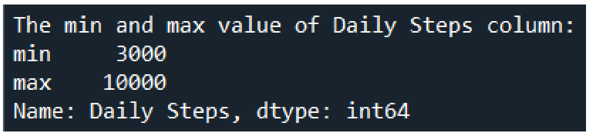
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Person ID                374 non-null    int64  
 1   Gender                   374 non-null    object 
 2   Age                      374 non-null    int64  
 3   Occupation               374 non-null    object 
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64  
 6   Physical Activity Level  374 non-null    int64  
 7   Stress Level             374 non-null    int64  
 8   BMI Category             374 non-null    object 
 9   Blood Pressure           374 non-null    object 
 10  Heart Rate               374 non-null    int64  
 11  Daily Steps              374 non-null    int64  
 12  Sleep Disorder           155 non-null    object 
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
None
```

In exploring the minimum and maximum value of a quantitative column, the *df.agg()* function can be used, as an example:

print("\nThe min and max value of Daily Steps column:")

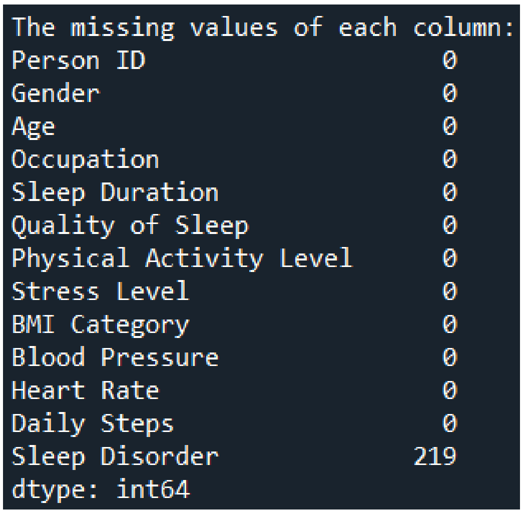
print(df['Daily Steps'].agg(['min', 'max']))

The ouput shows the minimum and maximum values of the Daily Steps column:



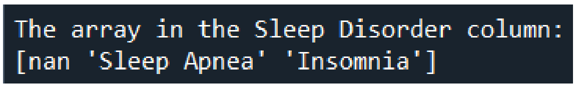
```
The min and max value of Daily Steps column:
min     3000
max    10000
Name: Daily Steps, dtype: int64
```

Next, data wrangling techniques imply finding and handling missing, incorrect, and invalid data. The first step is identifying the missing data by using the df.isnull().sum() function to check the dataset for null values. The Sleep Disorder column shows the presence of null values. Thus, to detect missing or irregular data, the print(df['Sleep Disorder'].unique()) command is used to see the array of the column.

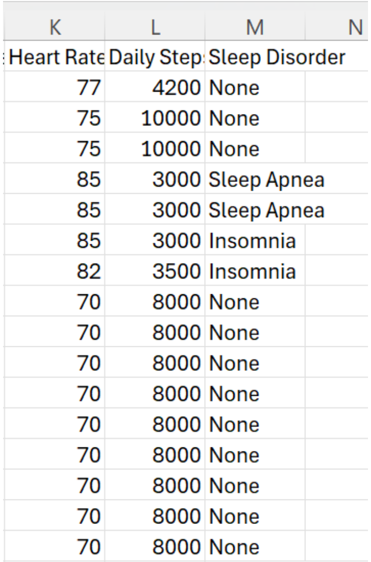


```
The missing values of each column:
Person ID                    0
Gender                       0
Age                          0
Occupation                   0
Sleep Duration               0
Quality of Sleep             0
Physical Activity Level      0
Stress Level                 0
BMI Category                 0
Blood Pressure               0
Heart Rate                   0
Daily Steps                  0
Sleep Disorder             219
dtype: int64
```

The array shows three values, which are NaN, Sleep Apnea, and Insomnia, but by checking manually the dataset, the NaN is actually referring to the ‘None’ string.



```
The array in the Sleep Disorder column:
[nan 'Sleep Apnea' 'Insomnia']
```

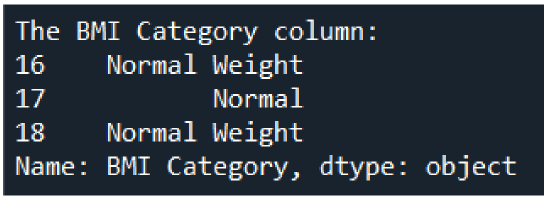


K	L	M	N
Heart Rate	Daily Step	Sleep Disorder	
77	4200	None	
75	10000	None	
75	10000	None	
85	3000	Sleep Apnea	
85	3000	Sleep Apnea	
85	3000	Insomnia	
82	3500	Insomnia	
70	8000	None	
70	8000	None	
70	8000	None	
70	8000	None	
70	8000	None	
70	8000	None	
70	8000	None	
70	8000	None	
70	8000	None	

Thus, to standardize the representation of missing values of the Sleep Disorder column, the ‘None’ string is replaced with NaN to ease the work on calculation or cleaning using *df['Sleep Disorder'] = df['Sleep Disorder'].replace(['None'], np.nan)* command.

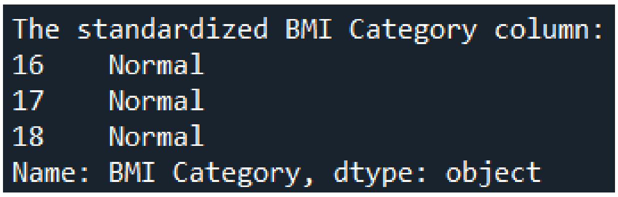
To add, the BMI Category column also shows inconsistent input from the users that required to be standardized, *df['BMI Category'] = df['BMI Category'].replace(['Normal Weight'], 'Normal')*, changing the ‘Normal Weight’ to ‘Normal’ to ensure all the data is presence for data visualization process.

Before standardizing:



```
The BMI Category column:
16    Normal Weight
17           Normal
18    Normal Weight
Name: BMI Category, dtype: object
```

After standardizing:



```
The standardized BMI Category column:
16    Normal
17    Normal
18    Normal
Name: BMI Category, dtype: object
```

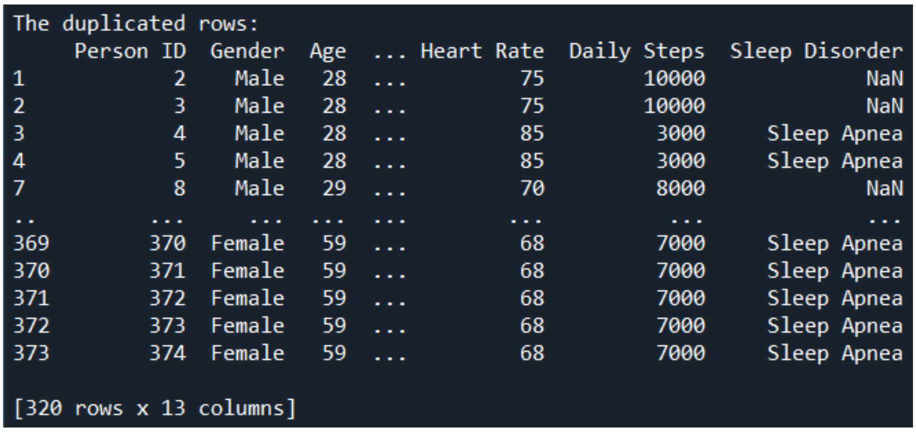
Next, we need to identify if any duplication data requires to be deleted using the *df.duplicated()* function as below:

duplicates = df.duplicated(subset=['Gender', 'Age', 'Occupation', 'Sleep Duration',
'Quality of Sleep', 'Physical Activity Level',
'Stress Level', 'BMI Category', 'Blood Pressure',
'Heart Rate', 'Daily Steps', 'Sleep Disorder'], keep=False)

Then, the duplication data is printed to see the list of duplicated rows.

print("\nThe duplicated rows:")
print(df[duplicates])

The output shows the dataset does have duplicate rows:



```
The duplicated rows:
     Person ID  Gender  Age  ... Heart Rate  Daily Steps  Sleep Disorder
1            2    Male   28  ...         75        10000             NaN
2            3    Male   28  ...         75        10000             NaN
3            4    Male   28  ...         85         3000     Sleep Apnea
4            5    Male   28  ...         85         3000     Sleep Apnea
7            8    Male   29  ...         70         8000             NaN
..         ...     ...  ...  ...        ...          ...             ...
369        370  Female   59  ...         68         7000     Sleep Apnea
370        371  Female   59  ...         68         7000     Sleep Apnea
371        372  Female   59  ...         68         7000     Sleep Apnea
372        373  Female   59  ...         68         7000     Sleep Apnea
373        374  Female   59  ...         68         7000     Sleep Apnea

[320 rows x 13 columns]
```

The duplicates need to be dropped from the dataset using the command below, with creating a new DataFrame *'df_updated'* to store the data after removing the duplicate rows:

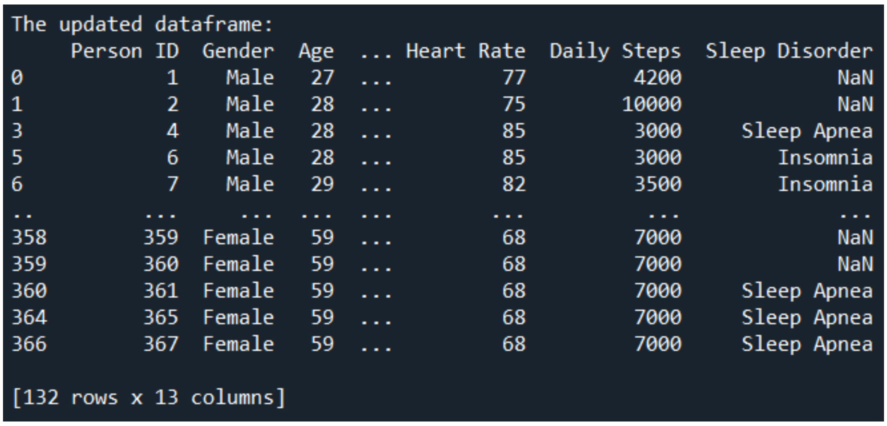
df_updated = df.drop_duplicates(subset=['Gender', 'Age', 'Occupation', 'Sleep Duration',
'Quality of Sleep', 'Physical Activity Level',

'Stress Level', 'BMI Category', 'Blood Pressure',
'Heart Rate', 'Daily Steps', 'Sleep Disorder'], keep='first')

To see the updated DataFrame, print the *df_updated* DataFrame.

print('\nThe updated dataframe:')
print(df_updated)

The output shows the number of rows decreasing from 320 to 132, after deleting the duplicate rows:



```
The updated dataframe:
     Person ID  Gender  Age  ... Heart Rate  Daily Steps  Sleep Disorder
0            1    Male   27  ...         77         4200             NaN
1            2    Male   28  ...         75        10000             NaN
3            4    Male   28  ...         85         3000     Sleep Apnea
5            6    Male   28  ...         85         3000        Insomnia
6            7    Male   29  ...         82         3500        Insomnia
..         ...     ...  ...  ...        ...          ...             ...
358        359  Female   59  ...         68         7000             NaN
359        360  Female   59  ...         68         7000             NaN
360        361  Female   59  ...         68         7000     Sleep Apnea
364        365  Female   59  ...         68         7000     Sleep Apnea
366        367  Female   59  ...         68         7000     Sleep Apnea

[132 rows x 13 columns]
```

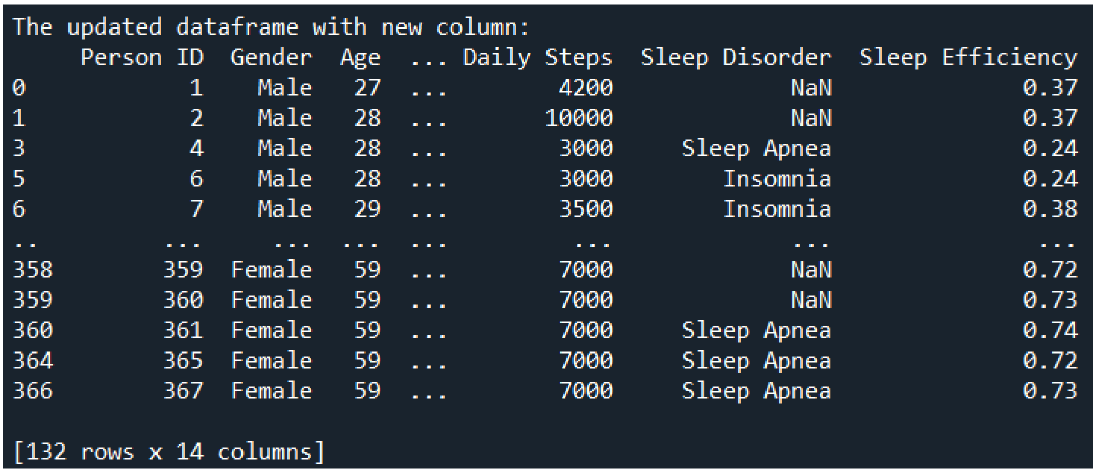
A new column has also been created to see the relationship between sleep efficiency with stress levels of individuals more precisely by multiplying the values from the ‘Sleep Duration’ and ‘Quality of Sleep’ columns, then dividing by 100. The product is then saved in 2 decimal places by using the *apply()* and *lambda* function.

*df_updated['Sleep Efficiency'] = ((df_updated['Sleep Duration'] * df_updated['Quality of Sleep']) / 100).apply(lambda x: '%.2f' % x)*

Finally, the new updated and standardized DataFrame is printed, including the new column added which is Sleep Efficiency.

print("\nThe updated dataframe with new column:")
print(df_updated)

The output shows the DataFrame columns increases from 13 to 14, adding the ‘Sleep Efficiency’ column:



```
The updated dataframe with new column:
     Person ID  Gender  Age  ... Daily Steps  Sleep Disorder  Sleep Efficiency
0            1    Male   27  ...        4200             NaN              0.37
1            2    Male   28  ...       10000             NaN              0.37
3            4    Male   28  ...        3000     Sleep Apnea              0.24
5            6    Male   28  ...        3000        Insomnia              0.24
6            7    Male   29  ...        3500        Insomnia              0.38
..         ...     ...  ...  ...         ...             ...               ...
358        359  Female   59  ...        7000             NaN              0.72
359        360  Female   59  ...        7000             NaN              0.73
360        361  Female   59  ...        7000     Sleep Apnea              0.74
364        365  Female   59  ...        7000     Sleep Apnea              0.72
366        367  Female   59  ...        7000     Sleep Apnea              0.73

[132 rows x 14 columns]
```

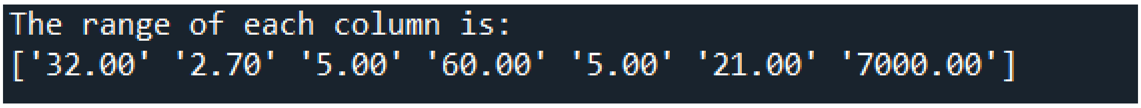
Exploratory Data Analysis

In this section, a few calculations were done for columns containing numerical values to see the dataset clearer. The columns are defined as below:

```
columns = ["Age","Sleep Duration", "Quality of Sleep", "Physical Activity Level",
                    "Stress Level", "Heart Rate", "Daily Steps"]
```

Range is calculated to study the distance between the maximum value and the minimum value in each numerical column. This value gives an idea of how far the data is dispersed. A larger range suggests a greater spread between the two values.

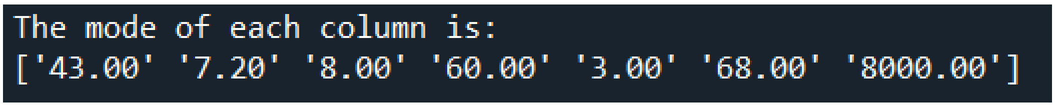
```
maximum = df[columns].max()
minimum = df[columns].min()
range= maximum - minimum
formatted_range = range[columns].apply(lambda x: "%.2f' % x)
print(f"The range of each column is:\n{formatted_range.values}\n")
```



```
The range of each column is:
['32.00' '2.70' '5.00' '60.00' '5.00' '21.00' '7000.00']
```

Mode shows the most frequent values in each column. It provides insight of the most typical value in the dataset, that can help us understand the dataset better. Mode is also helpful to detect potential outliers in our dataset.

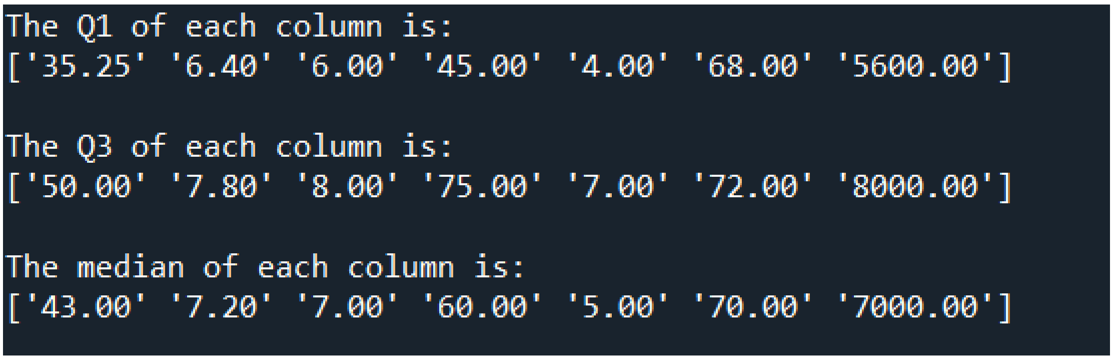
```
mode = df[columns].mode()
formatted_mode = mode[columns].apply(lambda x: "%.2f' % x)
print(f"The mode of each column is:\n{formatted_mode.values}\n")
```



```
The mode of each column is:
['43.00' '7.20' '8.00' '60.00' '3.00' '68.00' '8000.00']
```

Q1 and Q3 were counted to study the distribution of the values in each column. The first quantile, Q1 shows the value below 25% of the data. The third quantile, Q3 represents the upper value or the higher range of the data. Both quantiles are shown clearer in a box plot for outlier detection. Data that falls far from Q1 and Q3 are considered as outliers. Median is the value that falls in the middle of the data set. Median is not affected by outliers, which makes it reliable for central tendency representation for skewed distribution compared to average.

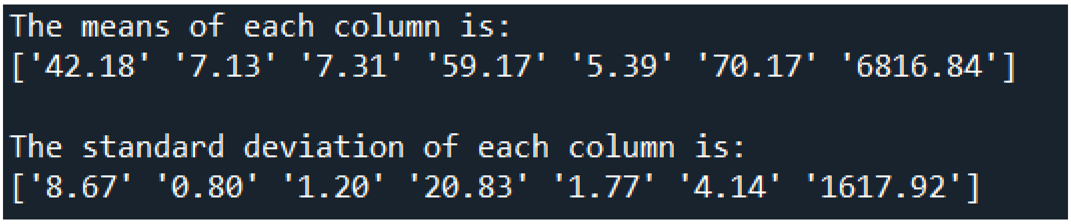
```
Q1=df[columns].quantile(0.25)
Q3=df[columns].quantile(0.75)
formatted_Q1 = Q1[columns].apply(lambda x: '%.2f' % x)
print(f"The Q1 of each column is:\n{formatted_Q1.values}\n")
formatted_Q3 = Q3[columns].apply(lambda x: '%.2f' % x)
print(f"The Q3 of each column is:\n{formatted_Q3.values}\n")
median = df[columns].median()
formatted_median = median[columns].apply(lambda x: '%.2f' % x)
print(f"The median of each column is:\n{formatted_median.values}\n")
```



```
The Q1 of each column is:
['35.25' '6.40' '6.00' '45.00' '4.00' '68.00' '5600.00']

The Q3 of each column is:
['50.00' '7.80' '8.00' '75.00' '7.00' '72.00' '8000.00']

The median of each column is:
['43.00' '7.20' '7.00' '60.00' '5.00' '70.00' '7000.00']
```

Mean is calculated by dividing the sum of all values with the number of values in a column. Mean summarizes the central location of the data. While median tells the center value and is not affected by outliers, mean is significantly affected by the extreme values in the data. Means are more meaningful in symmetric distribution and to check the skewness of the data. Standard deviation measures the amount of variation or the dispersion in a dataset. Standard deviation is related to mean as it indicates how far the values in a column are from the mean. A low standard deviation means values are close to the mean while a high standard deviation indicates the values in a column is spread out in a wider range from the mean.

```
means = df[columns].mean()
formatted_means = means[columns].apply(lambda x: '%.2f' % x)
print(f"The means of each column is:\n{formatted_means.values}\n")
std_dev=df[columns].std()
formatted_std_dev = std_dev[columns].apply(lambda x: '%.2f' % x)
print(f"The standard deviation of each column is:\n{formatted_std_dev.values}\n")
```

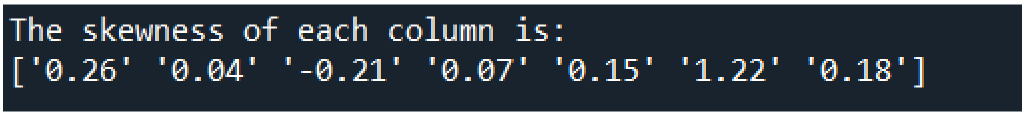


```
The means of each column is:
['42.18' '7.13' '7.31' '59.17' '5.39' '70.17' '6816.84']

The standard deviation of each column is:
['8.67' '0.80' '1.20' '20.83' '1.77' '4.14' '1617.92']
```

Another way to check for outliers without visual representation is by calculating the skewness of each column. As we can see, 'Age' column has a skewness of 0.23, which means it is slightly right skewed. This indicates more younger people are involved in this dataset. However, the skewness is significantly small, so we can say that the data is almost symmetric. 'Sleep Duration' column has a skewness of 0.03, which is also a small, slightly right-skewed data. It implies that almost all people in this study have a similar sleep duration. Next, the column 'Quality of Sleep' has a negative value of skewness, -0.21. This suggests that the data is slightly left-skewed. It might be affected by a few values that have an extremely low sleep quality that might affect the skewness.

After that, 'Physical Activity Level', 'Stress Level' and 'Daily Steps' column each has slightly right-skewed data. As the skewness is significantly low, we can conclude that it is almost symmetrical. These columns interpret that almost all of the subjects are living a similar and healthy lifestyle with standard physical activity level, low stress level, and having an average daily step. Last but not least, the 'Heart Rate' column has a skewness of 1.22, which is a bit more right-skewed compared to other columns. It suggests that most people have a relatively low heart rate, but very few of them have higher heart rates that is affecting the skewness of the data. We summarized all calculations done in a table for clearer interpretation.

```
skewness = df[columns].skew()
formatted_skewness = skewness[columns].apply(lambda x: '%.2f' % x)
print(f"The skewness of each column is:\n{formatted_skewness.values}\n")
```



```
The skewness of each column is:
['0.26' '0.04' '-0.21' '0.07' '0.15' '1.22' '0.18']
```

Column	Range	Q1	Median	Q3	Mode	Mean	Standard Deviation	Skewness
Age	32	35	43	50	43	42	8.67	0.26
Sleep Duration	2.70	6.40	7.20	7.80	7.20	7.13	0.80	0.04
Quality of Sleep	5.00	6.00	7.00	8.00	8.00	7.31	1.20	-0.21
Physical Activity Level	60.00	45.00	60.00	75.00	60.00	59.17	20.83	0.07
Stress Level	5.00	4.00	5.00	7.00	3.00	5.39	1.77	0.15
Heart Rate	21.00	68.00	70.00	72.00	68.00	70.17	4.14	1.22
Daily Steps	7000.0	5600.00	7000.0	8000.0	8000.0	6816.84	1617.92	0.18

Table 1 shows a summary of all calculations done in EDA.

Data Visualisation

- Explore distributions of numeric columns using histograms, etc.
- Explore relationship between columns using scatter plots, bar charts, etc.
- Explain your investigation process with your selected graphs.

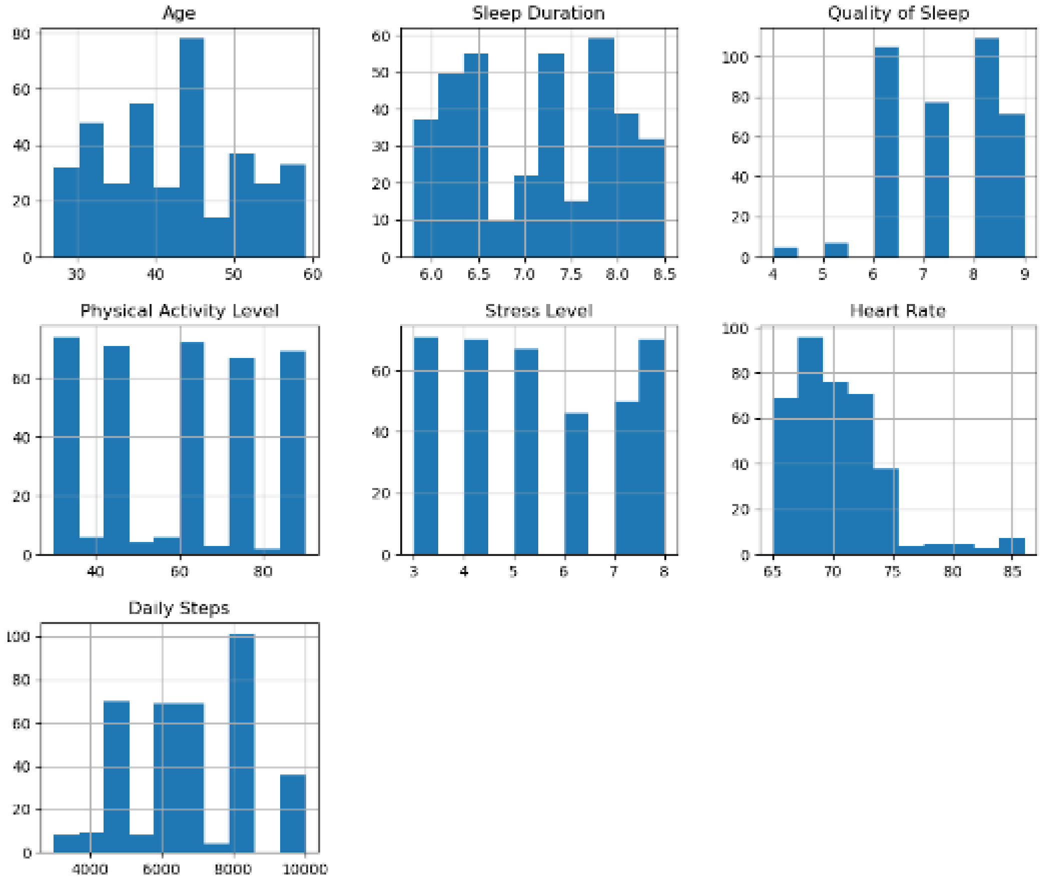
Explore distributions of numeric columns using histograms, etc.

Histograms

Histograms summarise data with descriptive statistics where the vertical axis represents the frequency while the horizontal axis represents the response variable.

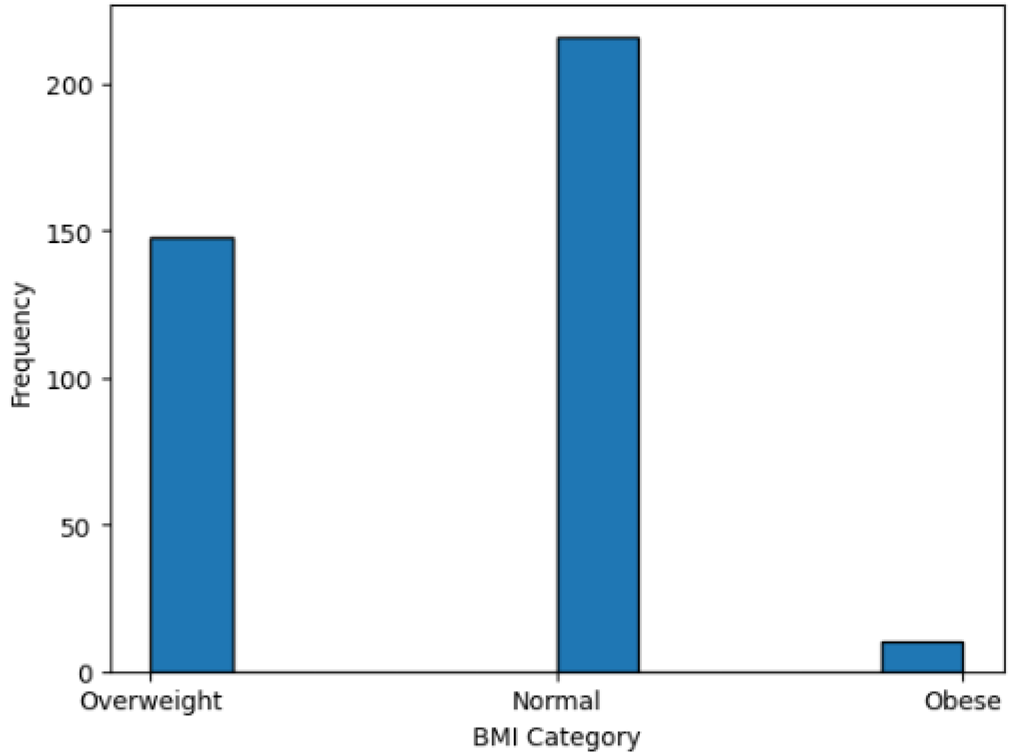
From the histogram below, we can observe that Age appears normally distributed with a bell-shaped curve and an almost symmetric graph. Most individuals in the dataset are clustered around a central age of 45, with fewer individuals at younger and older ages. Sleep Duration and Stress Level is a bimodal histogram because most of their data reside at two peaks. Physical Activity Level shows no significant skew. Heart Rate is right-skewed indicating a smaller individuals have higher heart rates while Quality of Sleep and Daily Steps are slightly left-skewed.

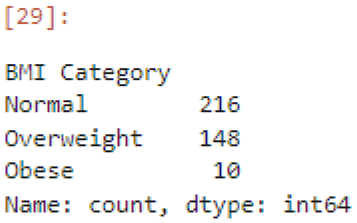
```
df[['Age', 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level',

  'Heart Rate', 'Daily Steps']].hist(figsize=(12, 10))

display()
```



Bar Plot

```
plt.hist(df['BMI Category'], histtype='bar', ec='black')

plt.xlabel('BMI Category')

plt.ylabel('Frequency')

plt.show()

df['BMI Category'].value_counts()
```

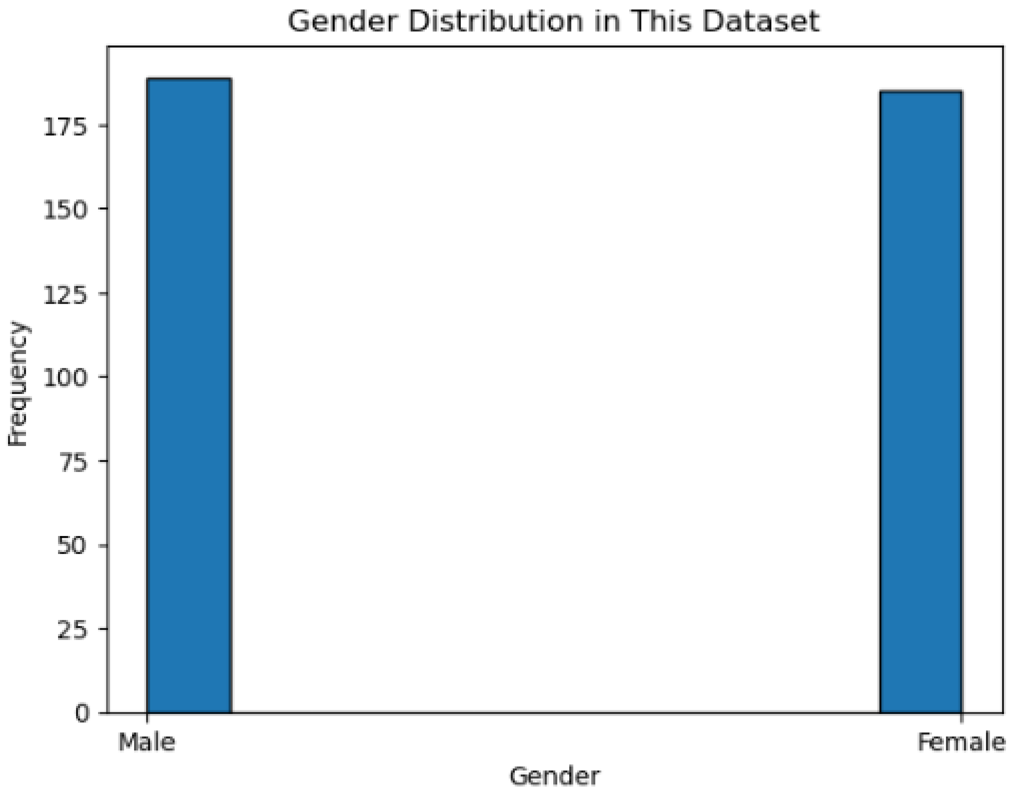


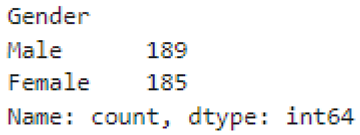


```
[29]:

BMI Category
Normal        216
Overweight    148
Obese          10
Name: count, dtype: int64
```

This bar plot shows the distribution of BMI categories in the dataset. The x-axis represents the BMI categories which consist of Normal, Overweight and Obese, and the y-axis represents the frequency or number of individuals falling under each category. Majority of the individuals have a BMI under the "Normal" category with a frequency of 216, followed by "Overweight" with 148 individuals and the "Obese" category having the lowest frequency with around 10 individuals.

```
plt.hist(df['Gender'], histtype='bar', ec='black')

plt.xlabel('Gender')

plt.ylabel('Frequency')

plt.title('Gender Distribution in This Dataset')
plt.show()

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



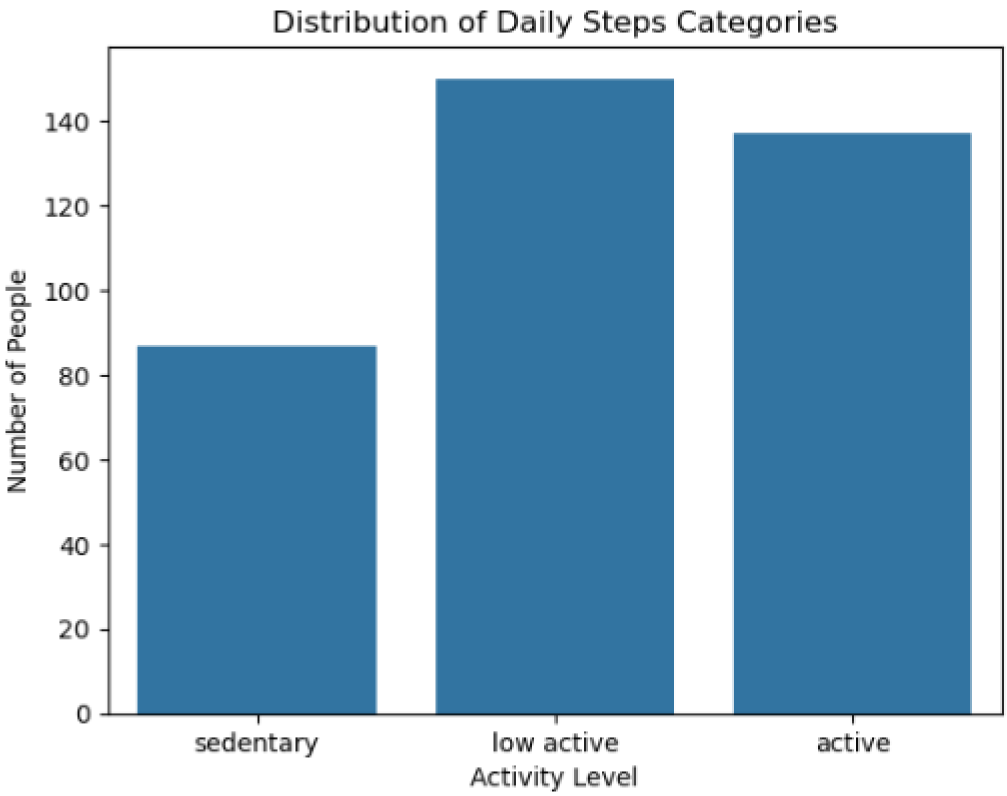


```
Gender
Male      189
Female    185
Name: count, dtype: int64
```

This is a bar plot representing the distribution of gender in this dataset. It is observed that both genders are almost uniformly distributed.

Binning Daily Steps and Presenting Using Bar Plot

```
bins = [0, 5000, 7500, 10000]

labels = ['sedentary', 'low active', 'active']

df['Steps Binned']=pd.cut(df['Daily Steps'], bins=bins, labels=labels)


sns.countplot(x='Steps Binned', data=df)

plt.xlabel('Activity Level')

plt.ylabel('Number of People')

plt.title('Distribution of Daily Steps Categories')

plt.show()
```



To categorise the daily steps, we created bins to represent the intervals by defining the bin edges as [0, 5000, 7500, 10000]. This represents four bins which are 0-5000, 5000-7500, and 7500-10000 representing sedentary, low active and active categories respectively. The cut() function was applied to the daily steps column and the results were stored in a new column called Steps Binned.

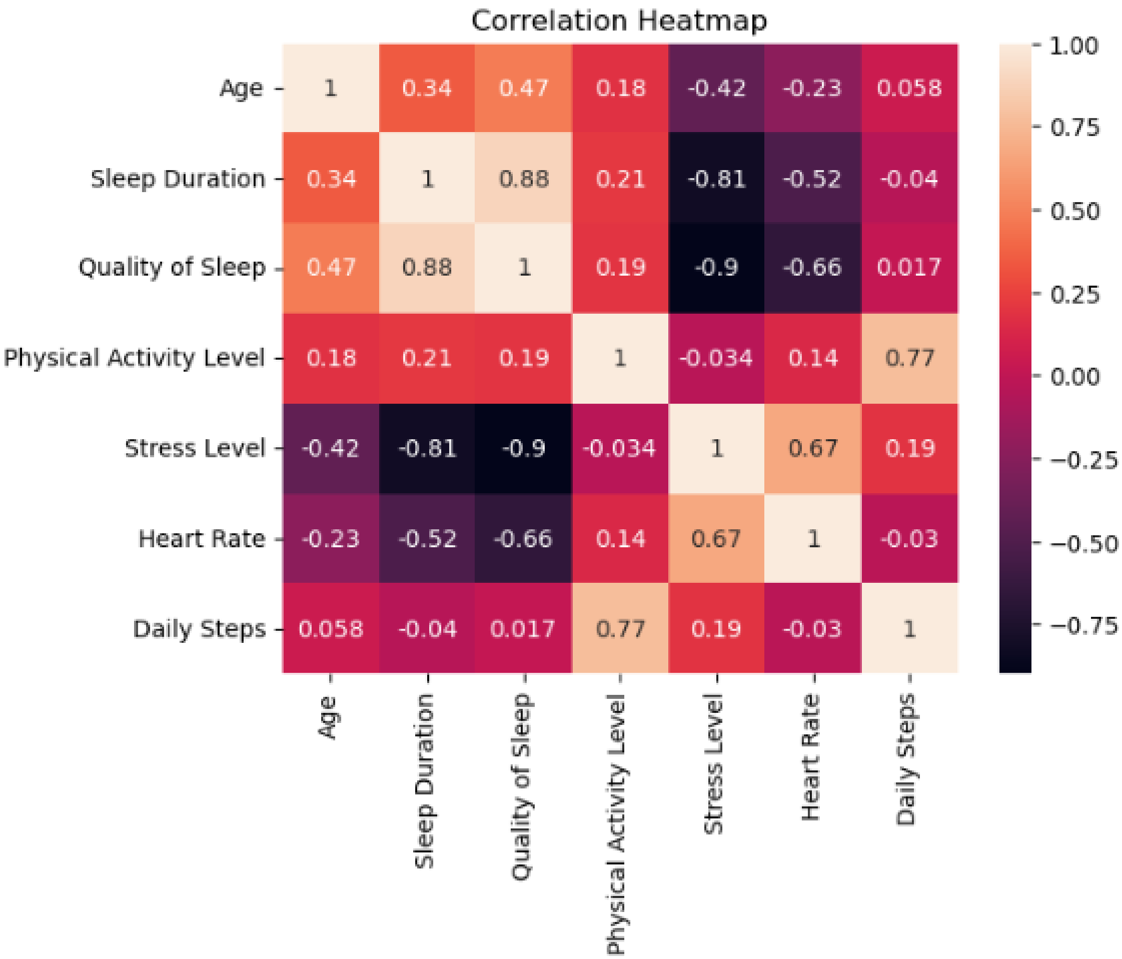
In this case, box plot is used to represent and compare the distributions across multiple categories. Box plot also provide a concise summary of data so that the key characteristics of the distribution can be clearly understand.

5.2 Explore relationship between columns using scatter plots, bar charts, etc.

```
!pip install matplotlib

import altair as alt
```

5.2.1 Heatmap

```
filtered = df.drop(['Person ID'], axis=1)

cor = filtered.select_dtypes(include='number').corr()

plt.figure(figure=(6,3))

sns.heatmap(cor, annot=True)

plt.title('Correlation Heatmap')

plt.show()
```

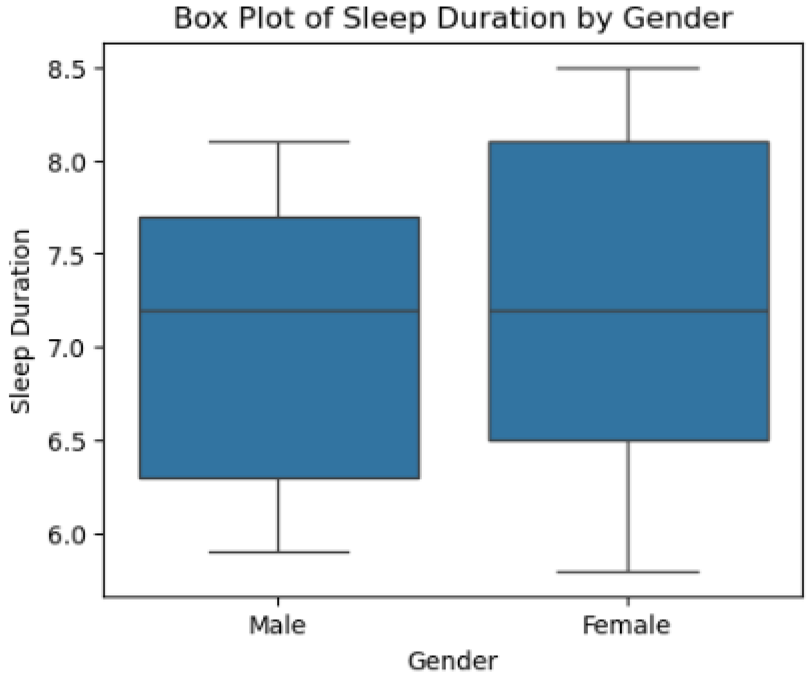


Heatmaps are used to show the relationships between two variables plotted on each axis. The relationship can be categorised as a positive correlation when a value closer to 1 indicates the relationship as one variable increases, the other also increases. A value closer to -1 indicates a strong negative correlation meaning the relationship as one variable increases, the other decreases. A value close to 0 suggests no relationship between the variables.

This heatmap shows the correlation between the numerical variables in the dataset. It can be observed that age has a weak positive correlation with the sleep metrics and a negative correlation with other factors such as stress level and heart rate. Besides, high stress has a significant negative impact on both sleep duration and sleep quality. There is a positive correlation between physical activity level and daily steps indicating more steps taken in a day correlates with higher physical activity levels. Higher heart rates are linked to poorer sleep and higher stress.

5.2.2 Box Plot

```
plt.figure(figsize=(5, 4))

sns.boxplot(x='Gender', y='Sleep Duration', data=df)

plt.title('Box Plot of Sleep Duration by Gender')

plt.show()
```

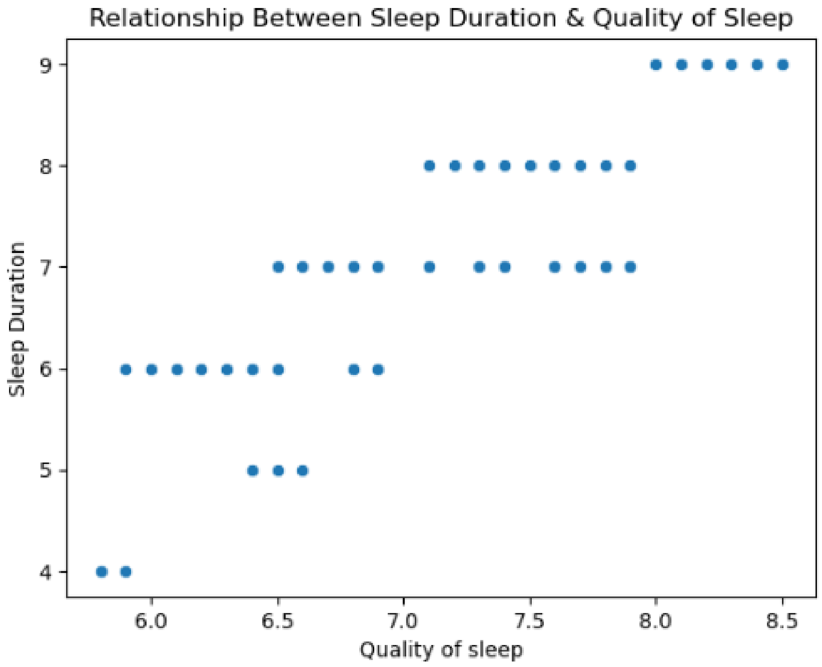


The box plot is useful to show the distribution of the data, and it can be used to represent the range of the data group with the minimum and maximum values. The box in the box plot represents the interquartile range of the dataset and the horizontal line represents the median.

The box plot above shows the sleep duration for male and female groups. The horizontal line inside the box represents the median sleep duration of both genders is around 7 to 7.5 hours. The interquartile range for females is wider than males, suggesting the sleep duration for females spread out a wider range of values than males.

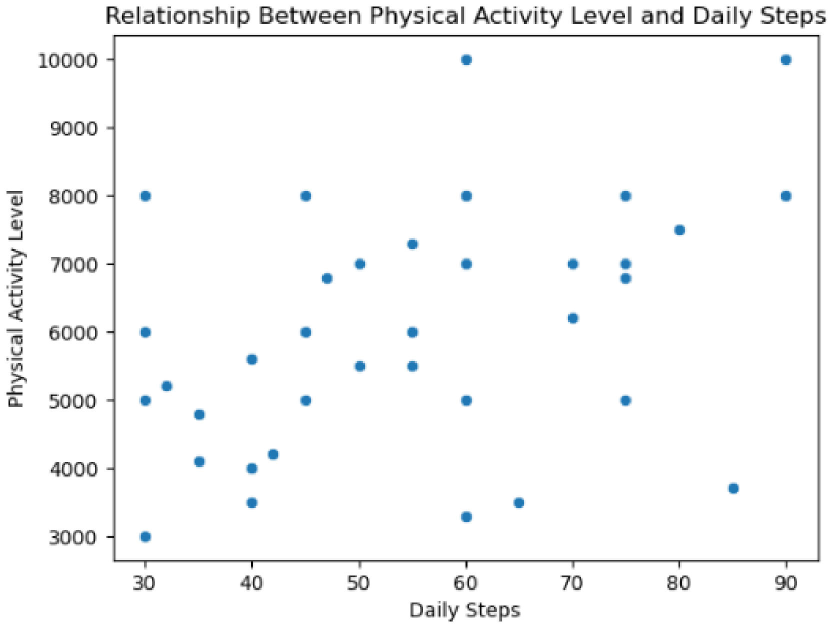
5.2.3 Scatter Plot

```
sns.scatterplot(x='Sleep Duration',y='Quality of Sleep', data=df)

plt.title('Relationship Between Sleep Duration & Quality of Sleep')

plt.ylabel('Sleep Duration')

plt.xlabel('Quality of sleep')

plt.show()
```



The scatter plot shows that when the sleep duration increases, the sleep quality also increases. It is used to show the relationships between two variables.

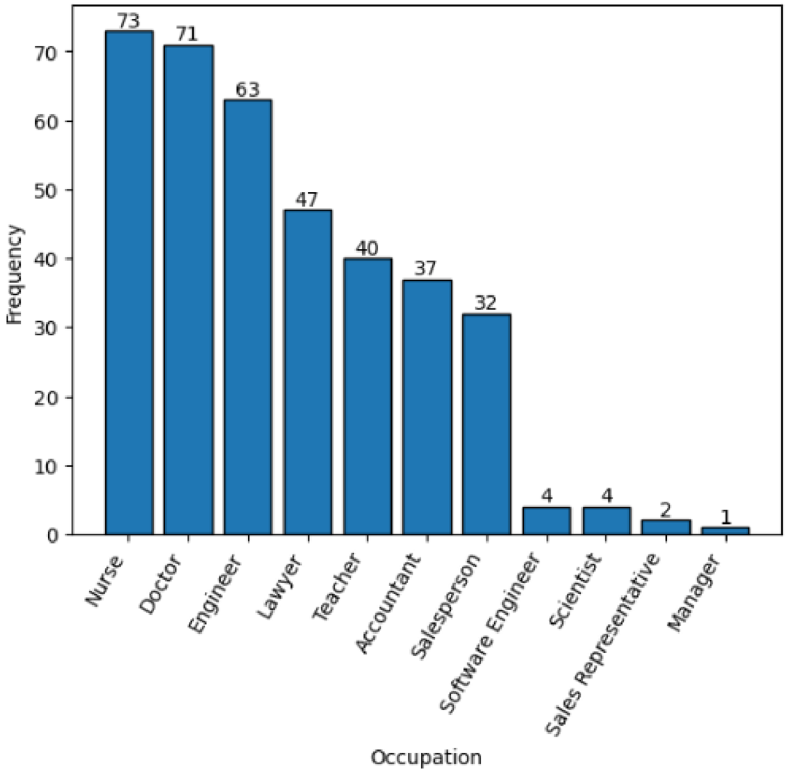
```
sns.scatterplot(x='Physical Activity Level',y='Daily Steps', data=df_updated)

plt.title('Relationship Between Physical Activity Level and Daily Steps')

plt.ylabel('Physical Activity Level')

plt.xlabel('Daily Steps')

plt.show()
```

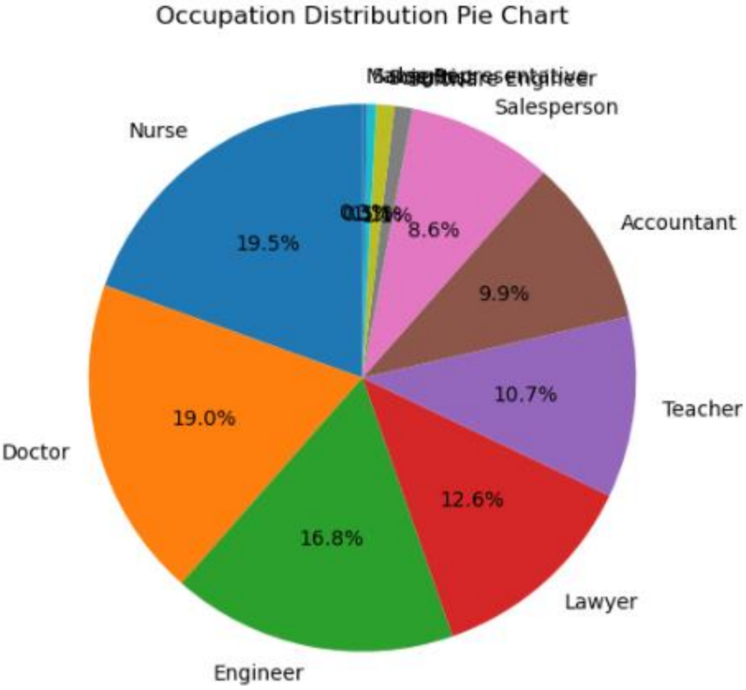


The scatter plot suggests that there is a weak positive correlation between physical activity level and daily steps which suggests there may be other factors influencing physical activity level. It is used to show the relationships between two variables.

5.2.4 Pie Chart and Count Plot

```
occupation_counts=df['Occupation'].value_counts()


plt.bar(occupation_counts.index, occupation_counts.values, edgecolor='black')
plt.xticks(rotation=60, ha='right')
plt.xlabel('Occupation')
plt.ylabel('Frequency')
for idx, value in enumerate(occupation_counts.values):
   plt.text(idx, value, str(value), ha='center', va='bottom', fontsize=10)
plt.show()


plt.figure(figsize=(6, 6))
plt.pie(occupation_counts, labels=occupation_counts.index, autopct='%1.1f%%', startangle=90)
plt.title('Occupation Distribution Pie Chart')
plt.show()
```


Frequency
0
10
20
30
40
50
60
70
73
71
63
47
40
37
32
4
4
2
1
Nurse
Doctor
Engineer
Lawyer
Teacher
Accountant
Salesperson
Software Engineer
Scientist
Sales Representative
Manager
Occupation


Occupation Distribution Pie Chart
Salesperson
Nurse
Accountant
Teacher
Doctor
Lawyer
Engineer
19.5%
8.6%
9.9%
10.7%
19.0%
12.6%
16.8%

The count plot and pie chart suggest the occupation distribution in this dataset which indicates that 19.5% of individuals are nurses, 19% are doctors, 16.8% are engineers, 12.6% are lawyers, 10.7% are teachers, followed by 9.9% accountants, 8.6% salesperson and others including software engineer, scientist, sales representatives and manager.

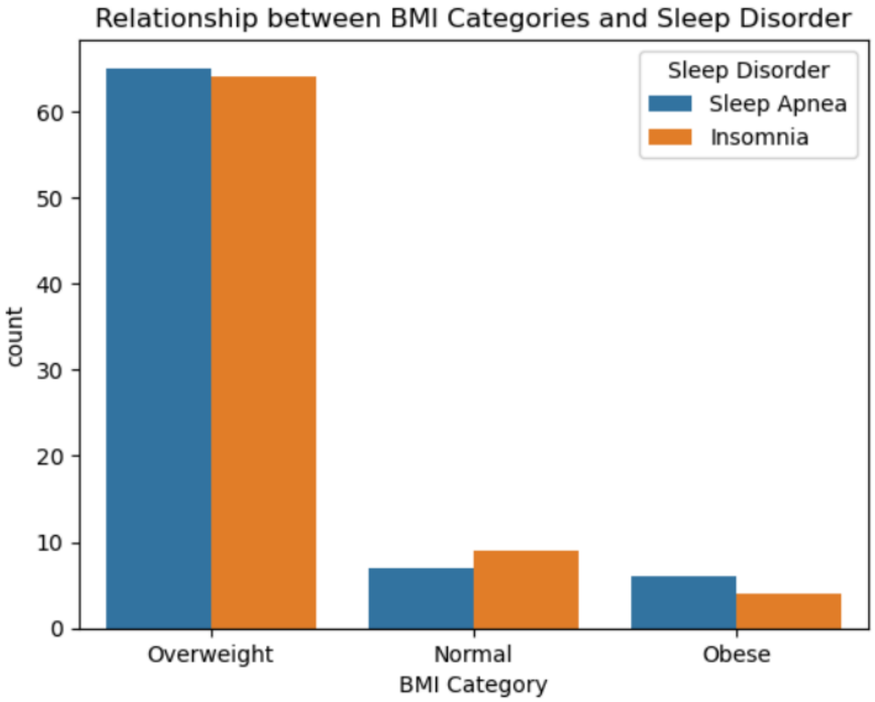
Discussion

In this section, the answers to the 4 proposed questions regarding the dataset are discussed and justified using data analytics method.

1. Does BMI influence ones' sleep disorder?

To answer this question, we are using a Count Plot to explore the relationship between BMI categories (Overweight, Normal, and Obese) and sleep disorder of individuals. Hence, based on the analysis made, people who are overweight are mostly diagnosed as having sleep disorders, which are Sleep Apnea and Insomnia, with the highest counts above 60 for both disorders. This clearly proves that there is a strong interdependence between being overweight and having sleep disorders.

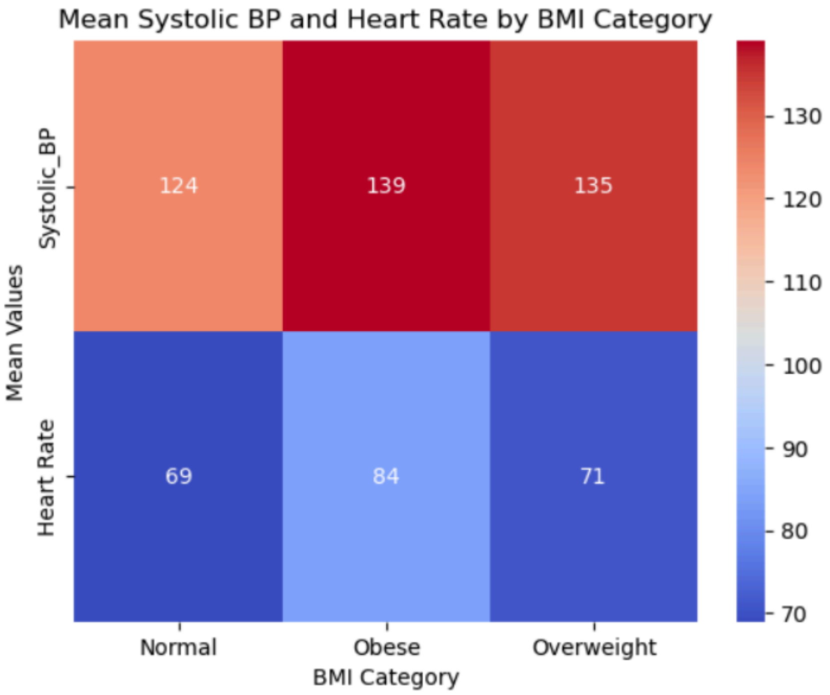
However, according to the graph below, people who are obese show the lowest count of sleep disorders among these 3 categories, with counts below 10 for both disorders. This concludes that BMI does not influence one's sleep disorder.



```
sns.countplot(x='BMI Category', hue='Sleep Disorder', data=df)
plt.title('Relationship between BMI Categories and Sleep Disorder')
plt.show()
```

2. How do BMI categories correlate with blood pressure and heart rate?

To find out the correlation between BMI categories with blood pressure and heart rate, we are using a Heatmap in this case. Thus, based on the Heatmap below, we can observe that as BMI increases, heart rate and systolic blood pressure also increases, indicating a positive correlation between BMI Categories with Heart Rate and Systolic_BP. Therefore, we can conclude that individuals with higher BMI, which are in the overweight and obese category, are most likely to have higher heart rates and blood pressure.

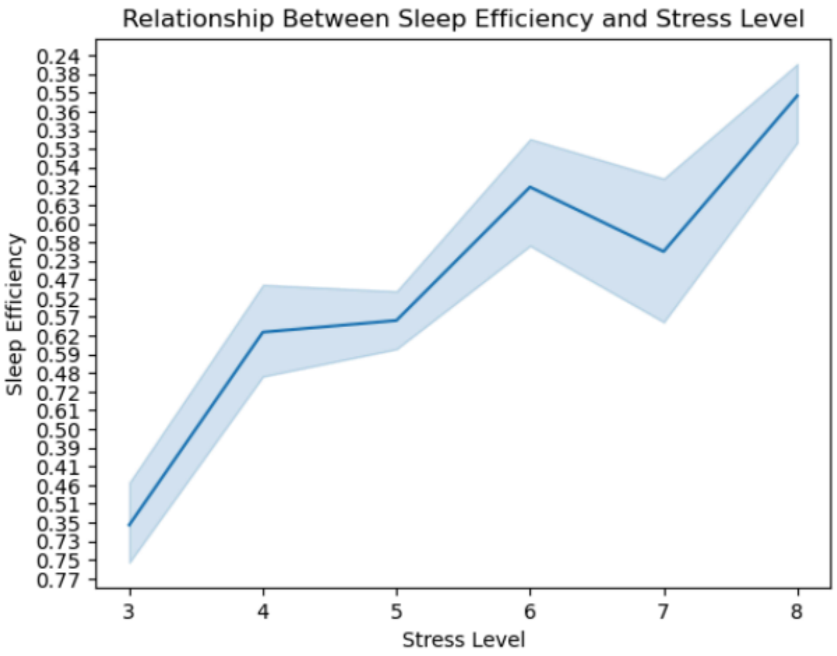


```
# Extract systolic value
df['Systolic_BP'] = df['Blood Pressure'].apply(lambda x: int(x.split('/')[0]))
heatmap_mean = df.groupby('BMI Category').agg({
    'Systolic_BP': 'mean',  # Mean of Systolic Blood Pressure
    'Heart Rate': 'mean'    # Mean of Heart Rate
}).reset_index()
print(heatmap_mean)

heatmap_mean['Systolic_BP'] = heatmap_mean['Systolic_BP'].round().astype(int)
heatmap_mean['Heart Rate'] = heatmap_mean['Heart Rate'].round().astype(int)
```

```
# heatmap
sns.heatmap(heatmap_mean.set_index('BMI Category').T, annot=True, fmt="d",
cmap='coolwarm', cbar=True)
plt.title('Mean Systolic BP and Heart Rate by BMI Category')
plt.xlabel('BMI Category')
plt.ylabel('Mean Values')
plt.show()
```

3. **Are there any trends or relationships between sleep efficiency and stress level?**

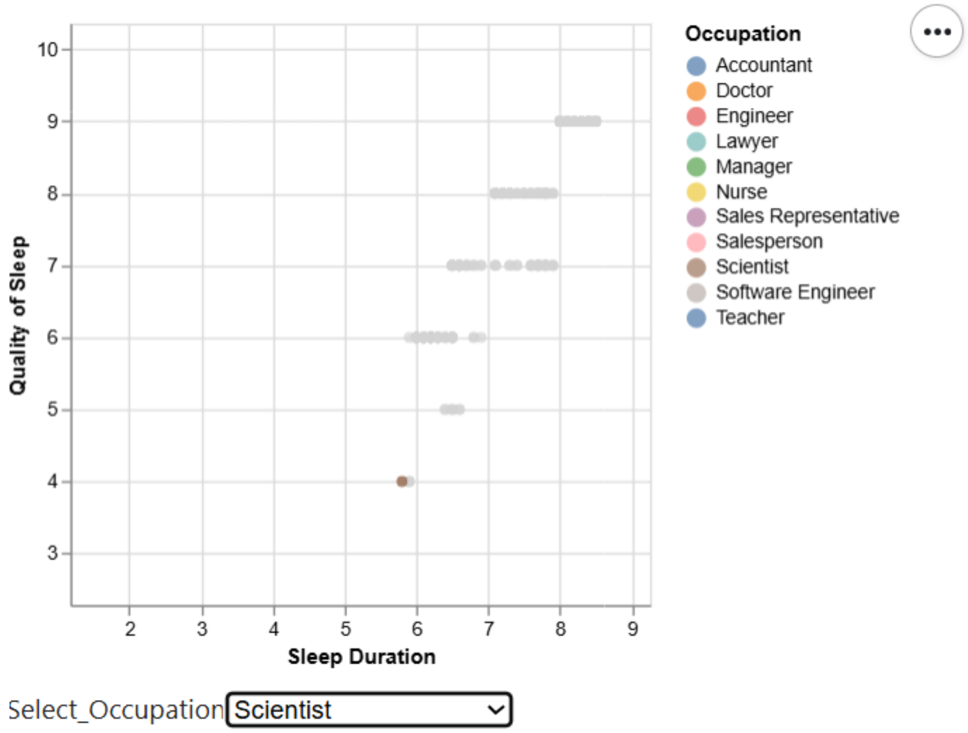
For this question, we are using a Line Chart to identify the trends or relationships between sleep efficiency and stress level. According to the Line Chart below, we observed that the stress level between level 3 until level 6 has a positive growth in sleep efficiency. Even though there is a fluctuation of sleep efficiency between stress levels 6 and 7, the entire line chart has a consistent increasing line, which results in steady growth of sleep efficiency over the stress levels. In conclusion, the sleep efficiency of an individual interdependence of their stress level.



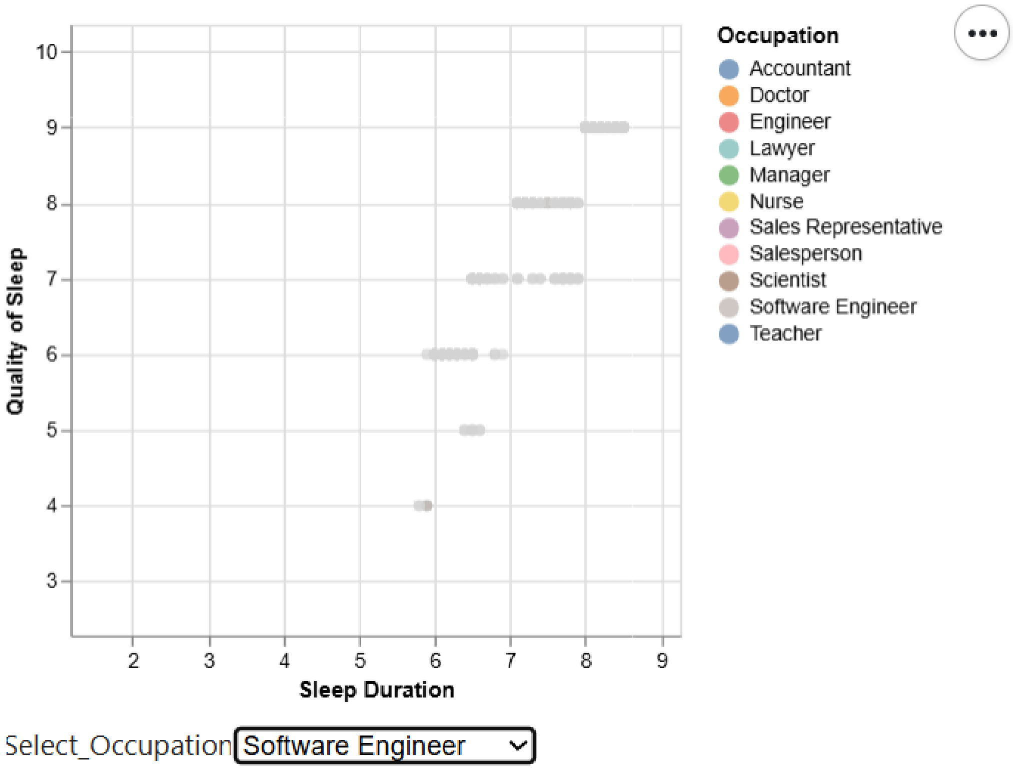
```
sns.lineplot(x='Stress Level', y='Sleep Efficiency', data=df_updated)
plt.title('Relationship Between Sleep Efficiency and Stress Level')
plt.ylabel('Sleep Efficiency')
plt.xlabel('Stress Level')
plt.show()
```

4. Which occupations experience the highest and lowest quality of sleep?

To answer this question, we are using an Interactive Scatter Plot to identify which occupations experience the highest and lowest quality of sleep. Hence, based on the plot below, it shows that Scientist and Software Engineer has the lowest quality of sleep and shortest sleep duration compared to other occupations.

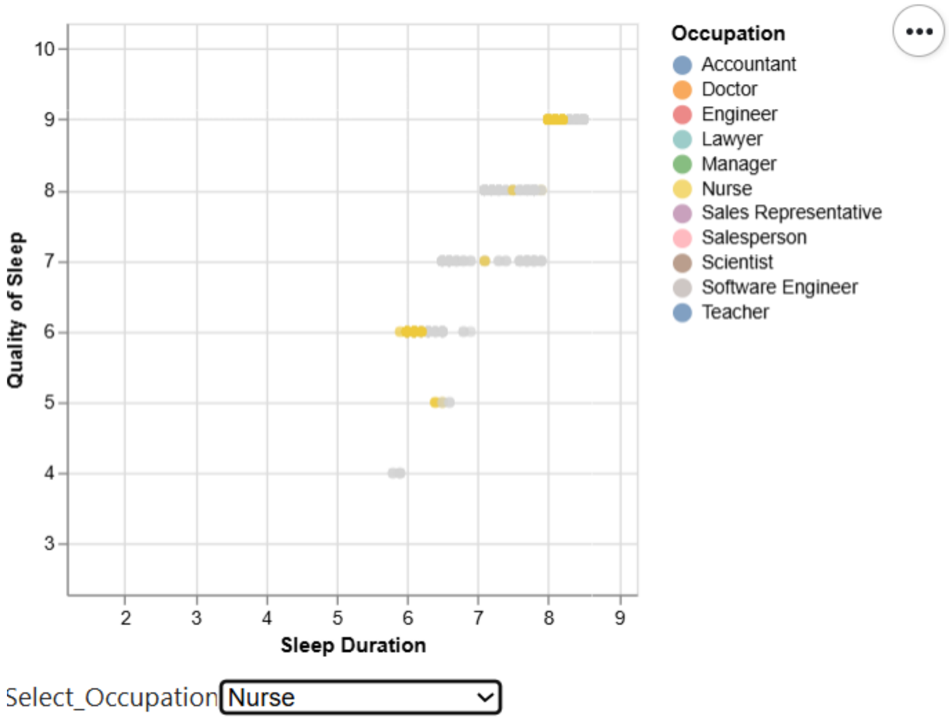


Select_Occupation Scientist

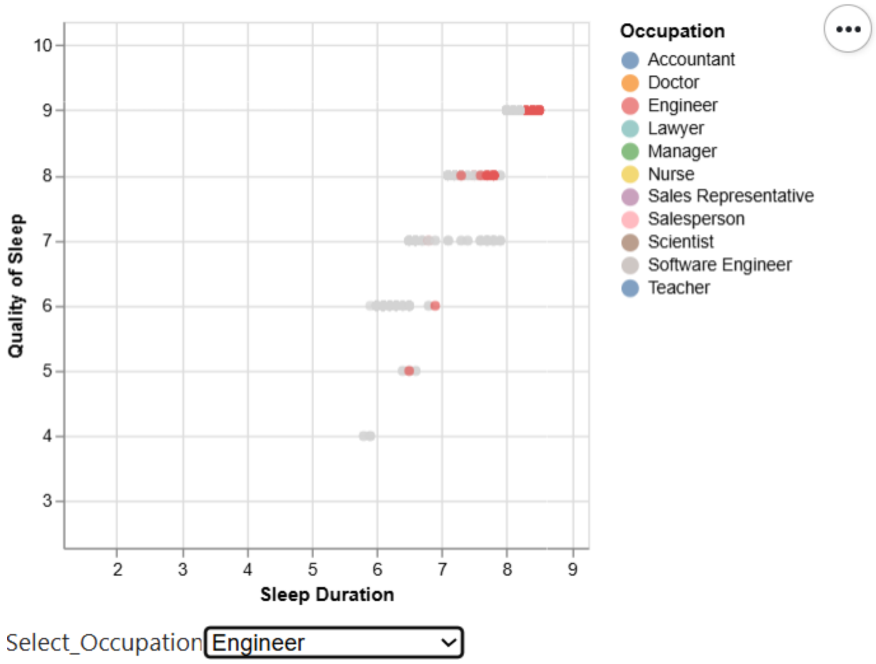


Select_Occupation Software Engineer

In contrast, Nurse and Engineer has the highest quality of sleep with longest sleep duration compared to others.



Select_Occupation Nurse



Select_Occupation Engineer

```
# dropdown menu for selecting an Occupation
input_dropdown = alt.binding_select(options=list(set(df.Occupation)))

selected_points = alt.selection_point(fields=['Occupation'], bind=input_dropdown,
name='Select')

color = alt.condition(selected_points, alt.Color('Occupation:N'), alt.value('lightgray'))

alt.Chart(df).mark_circle().encode(
  x='Sleep Duration',
  y='Quality of Sleep',
  color=color,
  tooltip=['Gender','Age', 'Occupation:N', 'Sleep Duration:Q', 'Quality of Sleep:Q', 'Sleep Disorder']
).add_params(
  selected_points
).interactive()
```

Conclusion

We are very grateful for having the opportunity to finish this project. We gained valuable insights through thorough research, analysis and practical application. The whole process of completing this project made us realize the importance of project planning, critical thinking, and communication skills between the group members. The challenges faced were overcome professionally by applying theoretical concepts and strategies and have prepared us for future assignments and upcoming endeavors. All in all, we would like to express our most heartfelt gratitude for the lecturer for assigning us this project and providing us with guidance and full support throughout this project. Their expertise and constructive feedback have been crucial for us and ensuring the completion of the project.

References

Bouchrika, I. (2024, 6 10). *Research.com*. Retrieved from 50 Current Student Stress Statistics: 2024 Data, Analysis & Predictions: https://research.com/education/student-stress-statistics

Bourne, P. A. (2011). Births, economic growth, mortality and murder in a developing country. *Health Vol.4 No.2*, 10.

C Joris Yzermans, G. A. (2005). Health problems of victims before and after disaster: a longitudinal study in general practice. *International Journal of Epidemiology*, 7.

Daharnis, Z. A. (2016). The compatibility student choice of university majoring: A preliminary studies. *GUIDENA: Journal of Guidance and Counseling*, 9.

Hanna Willa Dhany, S. F. (2023). Exploratory Data Analysis (EDA) methods for healthcare. *Journal of Intelligent Decision Support System (IDSS)*, 209-215.

Juanto Simangunsong, M. S. (2024). Mental disorder classification with exploratory data analysis (EDA). *Journal of Intelligent Decision Support System (IDSS)*, 210-217.

Juliana Antero, G. S.-F. (2018). A medal in the Olympics runs in the family: A cohort study of performance heritability in the Games history. *Frontiers in Physiology*, 10.

Kaweh Mansouri, F. A. (2013). Global rates of glaucoma surgery. *Graefes Arch Clin Exp Ophthalmol*, 7.

Kim, H.-J. (2022, September 10). *Wiley Online LIbrary*. Retrieved from The Analysis and Research on the Influence of Sports Industry Development on Economic Development: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/3329174

Liga Paura, I. A. (n.d.). Cause analysis of students' dropout rate in higher education study program. *Procedia - Social and Behavioral Sciences*, 5.

MFOWABO MAPHOSA, W. D. (2023). Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis. *IEEE Access*, 48977-48987.

Michiel van Dijk, Tom Morley, Marie Luise Rau, & Yashar Saghai. (2021). A meta-analysis of projected global food demand and population at risk of hunger for the period 2010-2050. *Nature Food*, 1-15.

Noviyanti T M Sagala, F. Y. (2022). Exploratory Data Analysis (EDA): A Study of Olympic. *SISTEMASI: Jurnal Sistem Informasi*, 578-587.

R. Indrakumari, T. Poongodi, & Soumya Ranjan Jena. (2020). Heart Disease Prediction using Exploratory Data Analysis. *Procedia Computer Science*, 130-139.

Roberto Barriga, Miquel Romero, Houcine Hassan, & David F. Nettleton. (2023). Energy Consumption Optimization of a Fluid Bed Dryer in Pharmaceutical Manufacturing Using EDA (Exploratory Data Analysis) . *Sensors*, 1-14.

Shaoyuan Lei, Rongshou Zeng, Siwei Zhang, Shaoming Wang, Ru Chen, Kexin Sun, . . . Wenqiang Wei. (2021). Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Communications*, 1-12.