Sleep Health and Lifestyle

Importance of Sleep and Health

- Sleep can have a positive impact on the human body including improving both physical and mental health
- Adults should get between 7-8 hours of sleep a night
- There are many factors that could impact a person's ability to get a meaningful amount of sleep

The Data

- The data set contained 13 columns with 400 observations
- The data set contained different variables that measure lifestyle factors, sleep metrics, cardiovascular health, and sleep disorder analysis
- We took the Blood Pressure and assigned values into Normal, Elevated, Hypertension Stage 1, Hypertension Stage 2, or Hypertension Crisis

Objective

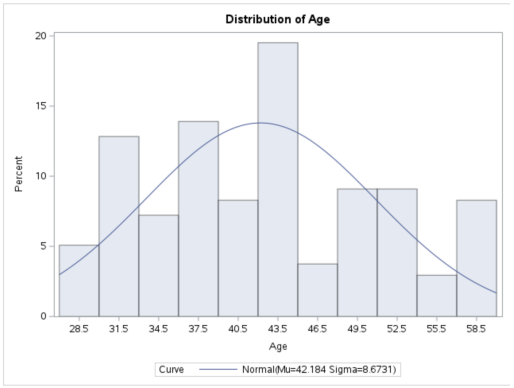
The project aims, through various statistic methods, to:

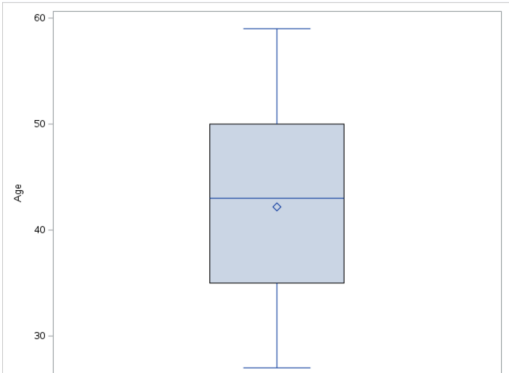
- Provide a comprehensive understanding of the key factors affecting sleep and health
- Identify potential areas of improvement for individuals based on lifestyle patterns
- Offer actionable insights for promoting healthier habits.

Overview of Univariate Statistics

Demographic Variables

- The average age was 42 years old
- The distribution of the ages is fairly normal
- Females made 49.47% of the observations with males at 50.53%

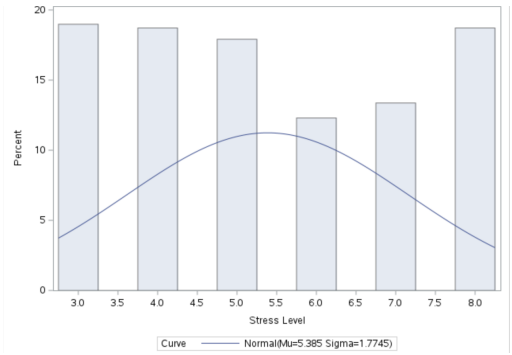


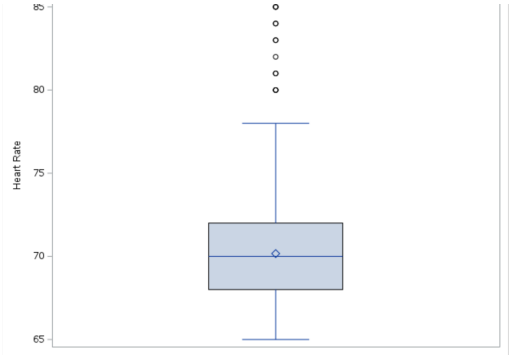


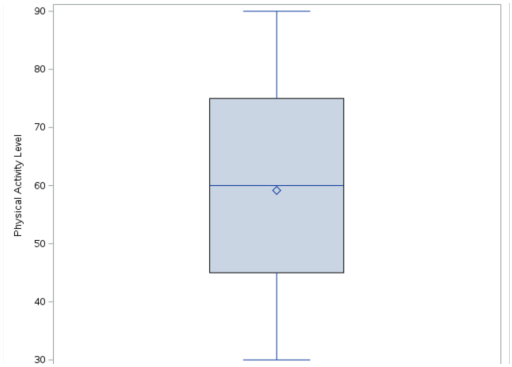
Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	185	49.47	185	49.47
Male	189	50.53	374	100.00

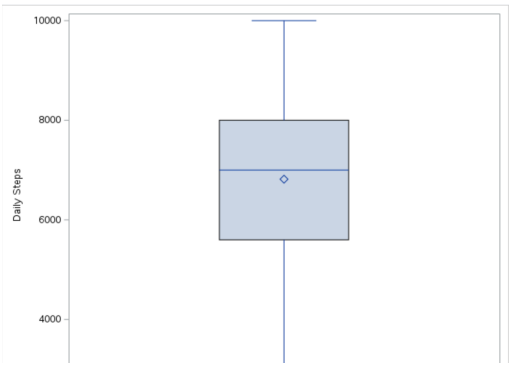
Lifestyle Variables

- The average physical activity of the patients was 59.17 and is fairly normal
- The average stress level of the patients was 5.39 and is normally distributed
- The average resting heart rate of the patients was 70.17 bpm and is slightly skewed right
- The average daily steps of the patients was 6,816.84 and is fairly normallay distributed



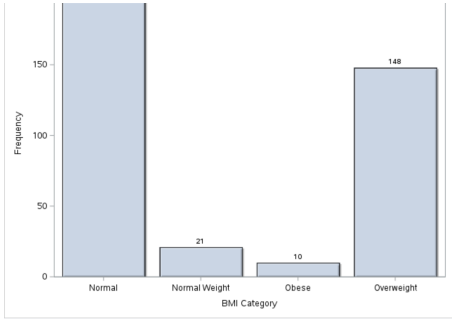


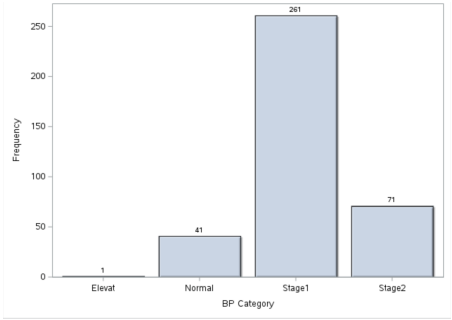


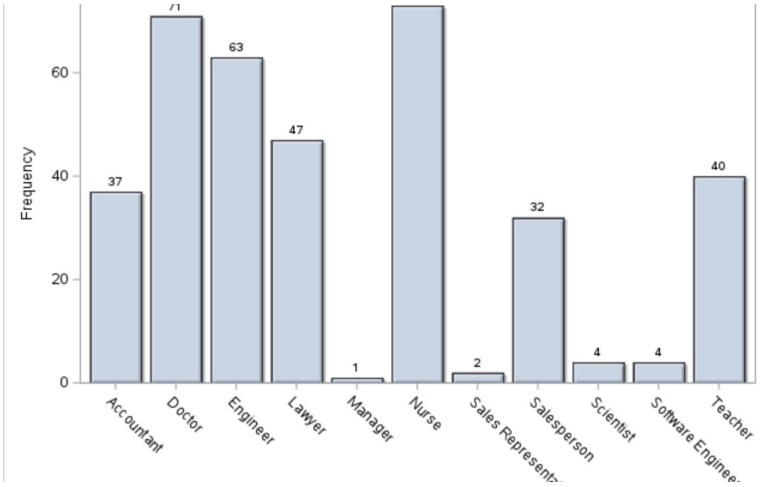


Lifestyle Variables cont.

- There were 11 different occupations with no more than 20% in any one category
- There were 4 different BMI categories with most of the patients being either normal or overweight
- Most of the patients had a blood pressure that was in Hypertension Stage 1

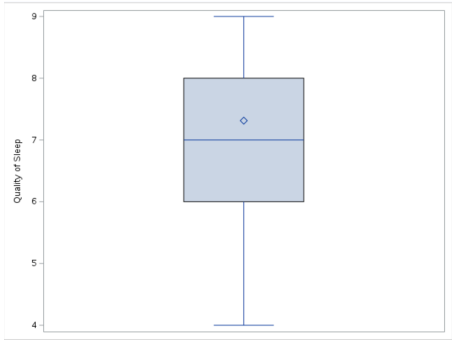


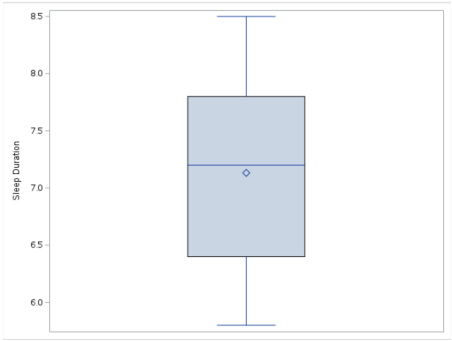


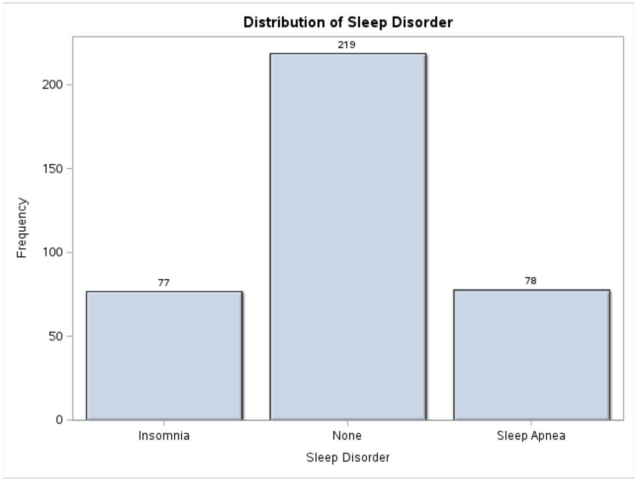


Sleep Related Variables

- Most of the patients did not have a sleep disorder
- The patients had an average of 7.13 hours of sleep per night and the distribution was normal
- The patients had an average sleep quality rating of 7.31 and was slightly skewed left







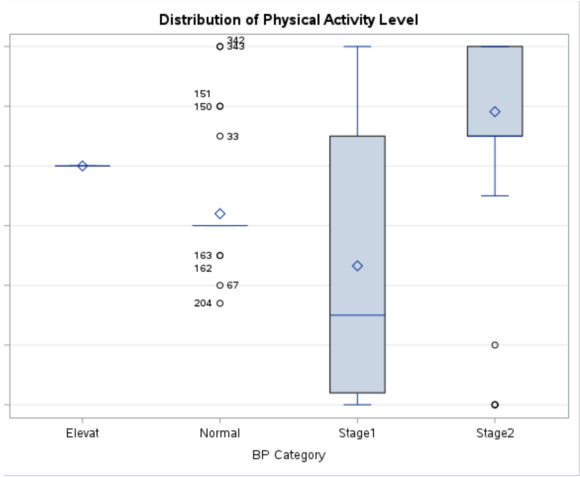
Bivariate Statistics

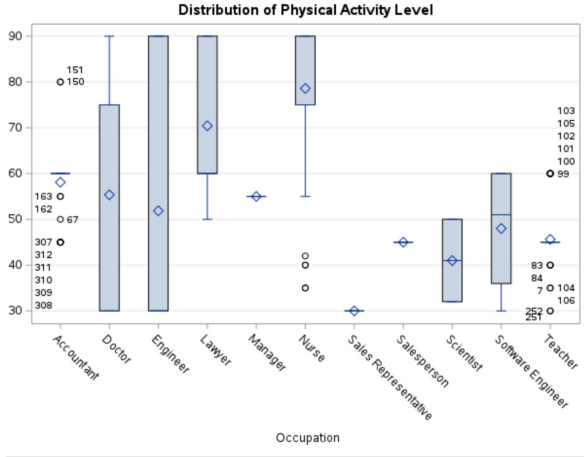
Physical Activity

- Stress Level did not have a statistically significant impact on Physical Activity. ($r= -0.034$ and $p=0.5$)
- Daily Steps had an impact of Physical Activity Level ($r=0.773$ and $p<0.0001$)
- Gender had a significant effect on Physical Activity ($p<0.0001$)
- Occupation had a significant effect on Physcial Activity ($p<0.0001$). Of the occupations, Nurses, lawyers, and teachers highlight differences in level
- The Blood Pressure category also showed a significant effect on Physical Activity ($p<0.0001$). Hypertensions Stage 2 patients showed to have higher activity levels

Physical Activity

Pearson Correlation Coefficients, N = 374 Prob > |r| under H0: Rho=0	Physical Activity Level
Age	0.17899 0.0005
Sleep Duration	0.21236 <.0001
Quality of Sleep	0.19290 0.0002
Stress Level	-0.03413 0.5105
Heart Rate	0.13697 0.0080
Daily Steps	0.77272 <.0001



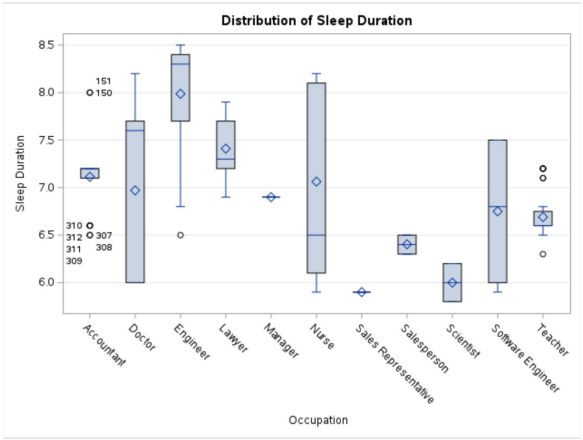


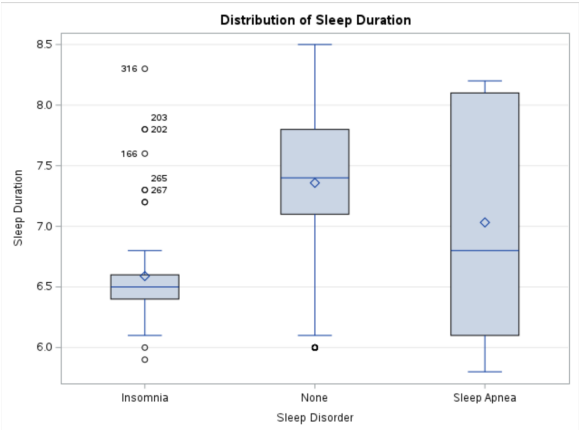
Sleep Duration

- Stress Level has a strong significant correlation relationship with Sleep Duration($r= -0.811$ and $p<0.0001$)
- Resting Heart Rate also significant correlation with Sleep Duration($r=-0.516$ and $p<0.0001$)
- Gender had a significant effect on Sleep Duration($p<0.0001$)
- Occupation had a significant effect on Sleep Duration($p<0.0001$). Of the occupations, engineers, salespersons, and scientists highlight significant differences in sleep duration
- Sleep Disorder had a significant effect on Sleep Duration($p=0.023$). There was a significant difference between the categories

Sleep Duration

Pearson Correlation Coefficients, N = 374 Prob > |r| under H0: Rho=0	
	Sleep Duration
Age	0.34471 <.0001
Physical Activity Level	0.21236 <.0001
Stress Level	-0.81102 <.0001
Heart Rate	-0.51645 <.0001
Quality of Sleep	0.88321 <.0001



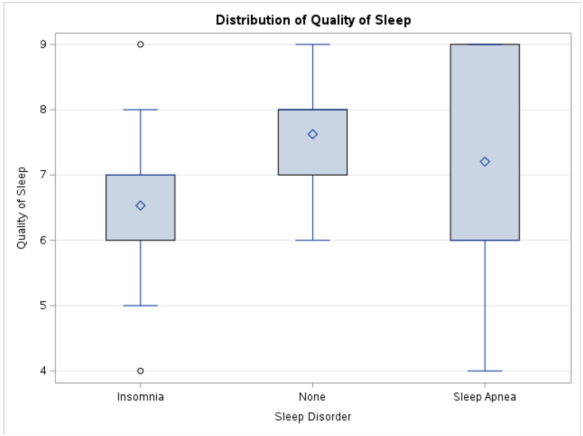


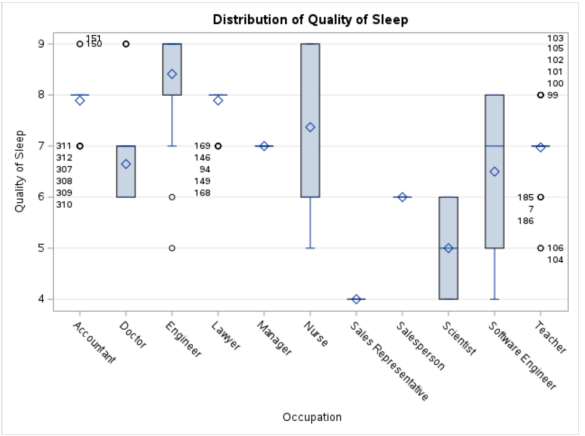
Sleep Quality

- Stress Level has a strong significant correlation relationship with Sleep Quality($r = -0.899$ and $p < 0.0001$)
- Age has a moderate significant correlation relationship with Sleep Quality($r = 0.474$ and $p < 0.0001$)
- Resting Heart Rate has a moderate significant correlation relationship with Sleep Quality($r = -0.660$ and $p < 0.0001$)
- Gender had a significant effect on Sleep Quality($p < 0.0001$)
- Occupation had a significant effect on Sleep Quality($p < 0.0001$). Of the occupations, engineers, lawyers and scientists highlight significant differences in sleep quality
- Sleep Disorder had a significant effect on Sleep Quality($p = 0.007$). There was a significant difference between the categories

Sleep Quality

Pearson Correlation Coefficients, N = 374 Prob > |r| under H0: Rho=0	
	Quality of Sleep
Age	0.47373 <.0001
Physical Activity Level	0.19290 0.0002
Stress Level	-0.89875 <.0001
Heart Rate	-0.65986 <.0001
Sleep Duration	0.88321 <.0001





Conclusions

- Occupation was a common factor across sleep and physical activity
- Sleep duration and Physical Activity did not have a strong correlation
- Stress level had a strong negative correlation with Sleep Duration and Sleep quality

Next Steps

- Do further experiments to see explore possible causes behind the correlations
- See what habits have the greatest effect on improved sleep quality and duration
- Add another variable related the urban vs. rural

References

- https://health.ucdavis.edu/blog/cultivating-health/better-sleep-why-its-important-for-your-health-and-tips-to-sleep-soundly/2023/03
- https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

SAS Code

Importing Our Dataset

```
/* Health Analysis */

/* Let's first import our dataset */

FILENAME REFFILE '/home/u63988465/Sleep_health_and_lifestyle_dataset.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=HEALTH;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=HEALTH; RUN;
```

Output:

Alphabetic List of Variables and Attributes

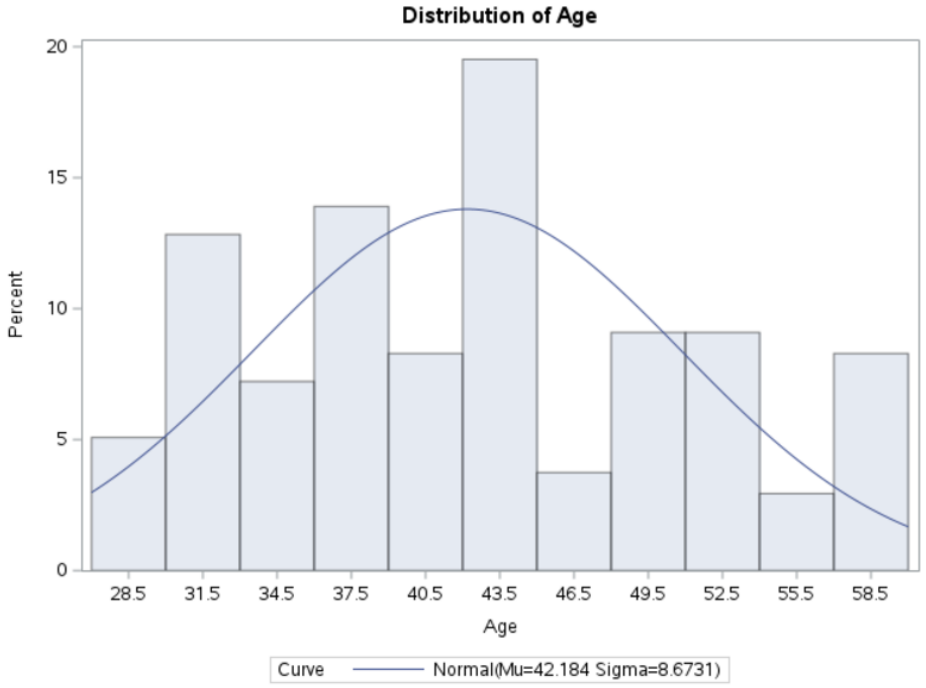
#	Variable	Type	Len	Format	Informat
3	Age	Num	8	BEST12.	BEST32.
9	BMI Category	Char	13	$13.	$13.
10	Blood Pressure	Char	6	$6.	$6.
12	Daily Steps	Num	8	BEST12.	BEST32.
2	Gender	Char	6	$6.	$6.
11	Heart Rate	Num	8	BEST12.	BEST32.
4	Occupation	Char	20	$20.	$20.
1	Person ID	Num	8	BEST12.	BEST32.
7	Physical Activity Level	Num	8	BEST12.	BEST32.
6	Quality of Sleep	Num	8	BEST12.	BEST32.
13	Sleep Disorder	Char	11	$11.	$11.
5	Sleep Duration	Num	8	BEST12.	BEST32.
8	Stress Level	Num	8	BEST12.	BEST32.

Obs	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
1	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
2	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea

Our Dataset contains 13 Features and 374 Observation

```sas
/* Now let's check the each variable individually and describe the data (shape, form, spread, outliers) */

/* We will only check for Numerical Variable */

/* 1. Age */
PROC UNIVARIATE DATA=HEALTH;
    VAR Age;
    HISTOGRAM / NORMAL;
RUN;
```

Output:



The UNIVARIATE Procedure
Variable: Age

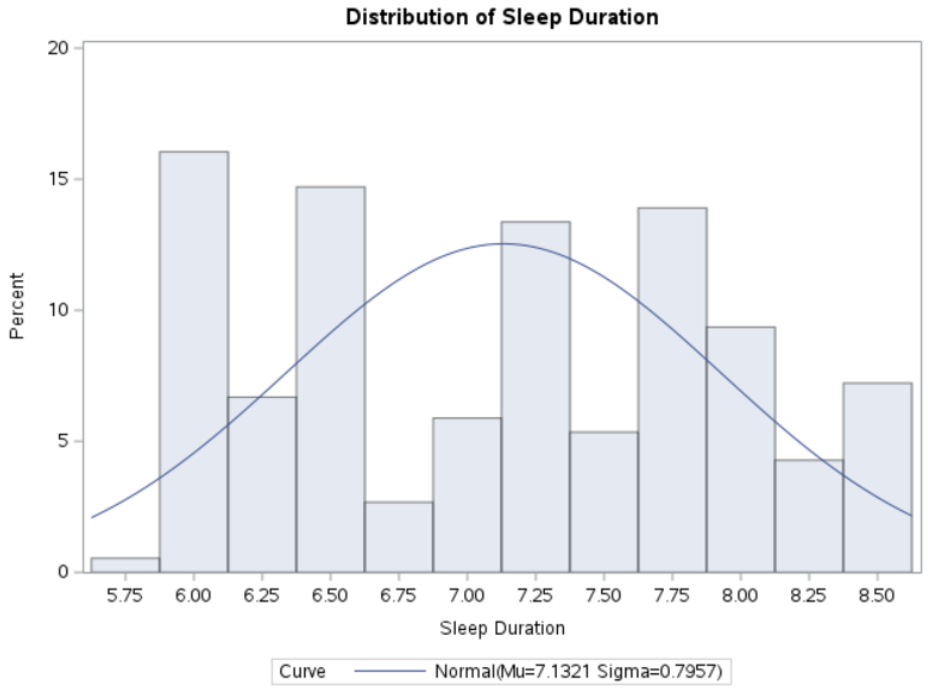
Moments			
N	374	Sum Weights	374
Mean	42.184492	Sum Observations	15777
Std Deviation	8.67313347	Variance	75.2232441
Skewness	0.25722214	Kurtosis	-0.9097795
Uncorrected SS	693603	Corrected SS	28058.2701
Coeff Variation	20.5600045	Std Error Mean	0.44847705

Basic Statistical Measures			
Location		Variability	
Mean	42.18449	Std Deviation	8.67313
Median	43.00000	Variance	75.22324
Mode	43.00000	Range	32.00000
		Interquartile Range	15.00000

The distribution of age appears to be roughly symmetric, with a mean of 42.18, a median of 43, and a mode of 43. The standard deviation of 8.67 indicates moderate variability, and the skewness of 0.26 suggests a slight positive skew in the data.

```sas
/* 2. Sleep Duration */
PROC UNIVARIATE DATA=HEALTH;
    VAR 'Sleep Duration'n;
    HISTOGRAM / NORMAL;
RUN;
```

Output:



The UNIVARIATE Procedure
Variable: Sleep Duration

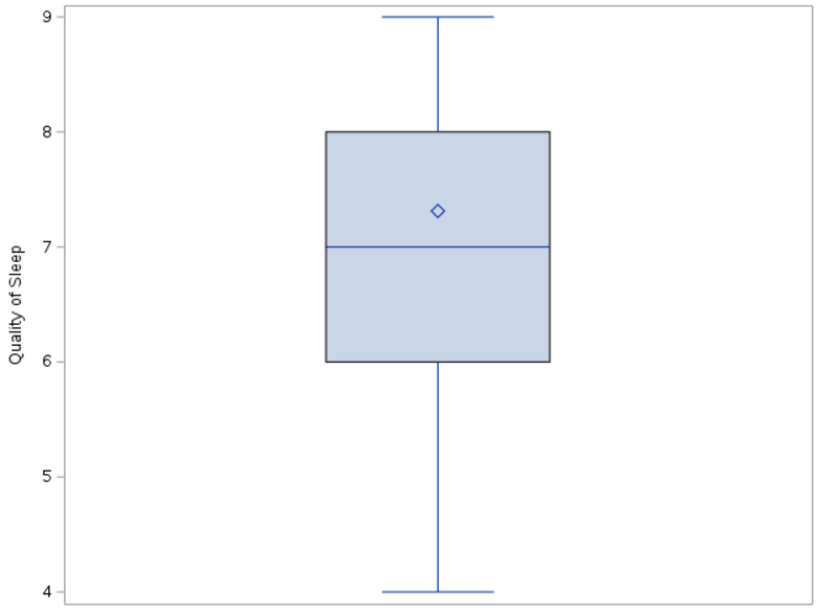
Moments			
N	374	Sum Weights	374
Mean	7.13208556	Sum Observations	2667.4
Std Deviation	0.79565673	Variance	0.63306963
Skewness	0.03755439	Kurtosis	-1.2865062
Uncorrected SS	19260.26	Corrected SS	236.134973
Coeff Variation	11.1560177	Std Error Mean	0.04114243

Basic Statistical Measures			
Location		Variability	
Mean	7.132086	Std Deviation	0.79566
Median	7.200000	Variance	0.63307
Mode	7.200000	Range	2.70000
		Interquartile Range	1.40000

The distribution of sleep duration is relatively symmetric, with a mean of 7.13 hours and a median and mode of 7.2 hours, indicating consistency in central tendency. The standard deviation is 0.80, showing low variability in the data, and the range is 2.7 hours. The skewness (0.037) and kurtosis (-1.287) suggest a nearly normal distribution with a slightly flatter peak. The interquartile range (1.4 hours) highlights the compact spread of middle values.

```sas
/* 3. Quality of Sleep */
PROC UNIVARIATE DATA=HEALTH;
    VAR 'Quality of Sleep'n;
    HISTOGRAM / NORMAL;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBOX 'Quality of Sleep'n/ NAME='Boxplot For Quality of Sleep';
RUN;
```

Output:



The UNIVARIATE Procedure
Variable: Quality of Sleep

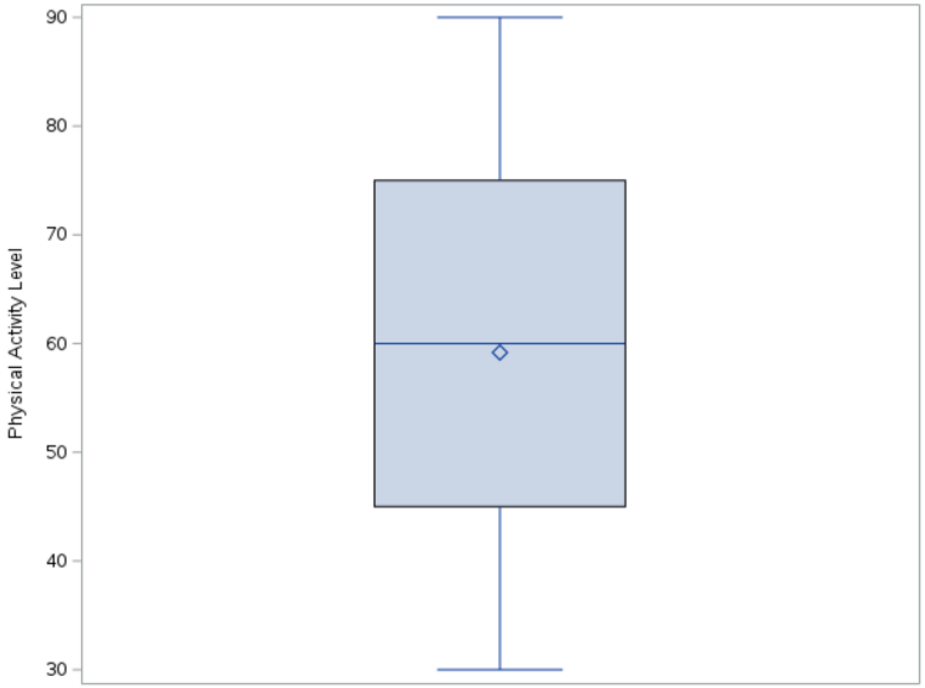
Moments			
N	374	Sum Weights	374
Mean	7.31283422	Sum Observations	2735
Std Deviation	1.19695592	Variance	1.43270347
Skewness	-0.2074476	Kurtosis	-0.7482755
Uncorrected SS	20535	Corrected SS	534.398396
Coeff Variation	16.3678799	Std Error Mean	0.06189312

Basic Statistical Measures			
Location		Variability	
Mean	7.312834	Std Deviation	1.19696
Median	7.000000	Variance	1.43270
Mode	8.000000	Range	5.00000
		Interquartile Range	2.00000

The quality of sleep data shows a mean of 7.31 and a median of 7, indicating a slight symmetry in distribution. The standard deviation of 1.20 reflects moderate variability, while the interquartile range of 2 suggests a relatively consistent middle spread. The box plot shows no significant outliers, and the mode of 8 highlights the most frequently reported sleep quality. The skewness (-0.21) and kurtosis (-0.75) indicate a slight negative skew and a flatter distribution than normal.

```sas
/* 4. Physical Activity Leel */
PROC UNIVARIATE DATA=HEALTH;
    VAR 'Physical Activity Level'n;
    HISTOGRAM / NORMAL;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBOX 'Physical Activity Level'n/ NAME='Boxplot For Physical Activity Level';
RUN;
```

Output:



The UNIVARIATE Procedure
Variable: Physical Activity Level

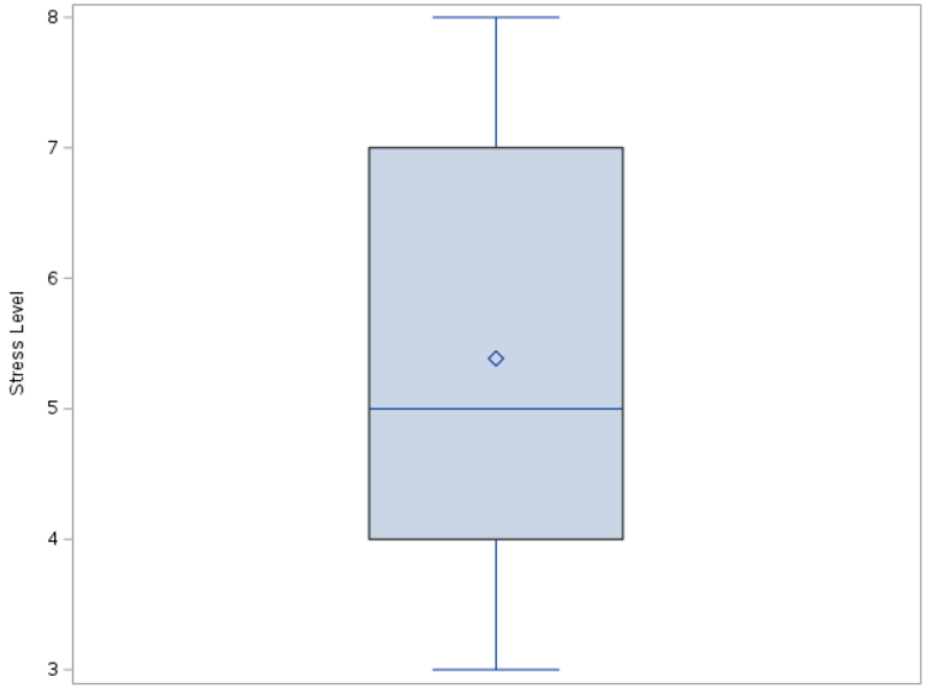
Moments			
N	374	Sum Weights	374
Mean	59.171123	Sum Observations	22130
Std Deviation	20.8308037	Variance	433.922381
Skewness	0.0744869	Kurtosis	-1.2660678
Uncorrected SS	1471310	Corrected SS	161853.048
Coeff Variation	35.2043406	Std Error Mean	1.07713521

Basic Statistical Measures			
Location		Variability	
Mean	59.17112	Std Deviation	20.83080
Median	60.00000	Variance	433.92238
Mode	60.00000	Range	60.00000
		Interquartile Range	30.00000

The physical activity level data shows a mean of 59.17 and a median of 60, indicating a near-symmetric distribution. The standard deviation of 20.83 reflects a moderate spread in activity levels, with an interquartile range of 30, capturing the variability in the central 50% of data. The skewness (0.07) and kurtosis (-1.27) suggest a nearly symmetric and flatter-than-normal distribution. The range of 60 indicates substantial variability in the physical activity levels among individuals.

```
/* 5. Stress Level */
PROC UNIVARIATE DATA=HEALTH;
    VAR 'Stress Level'n;
    HISTOGRAM / NORMAL;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBOX 'Stress Level'n/ NAME='Boxplot For Stress Level';
RUN;
```

Output:



The UNIVARIATE Procedure
Variable: Stress Level

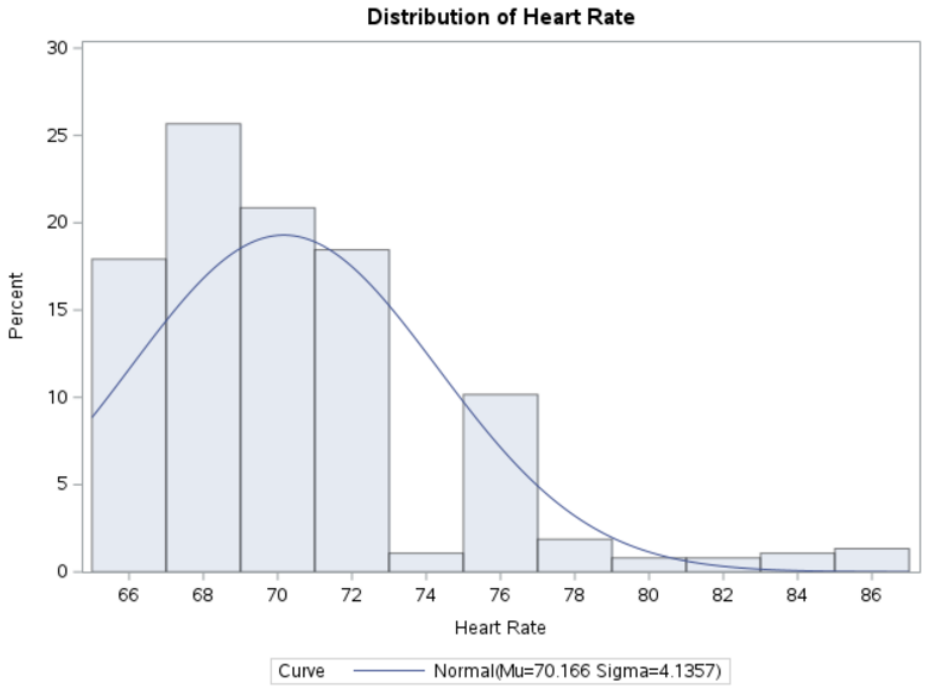
Moments			
N	374	Sum Weights	374
Mean	5.38502674	Sum Observations	2014
Std Deviation	1.77452644	Variance	3.1489441
Skewness	0.15432958	Kurtosis	-1.3273066
Uncorrected SS	12020	Corrected SS	1174.55615
Coeff Variation	32.9529737	Std Error Mean	0.09175858

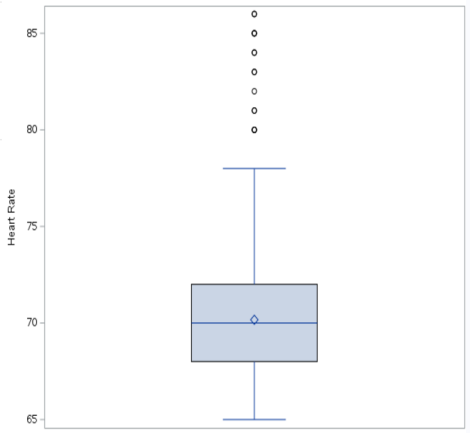
Basic Statistical Measures			
Location		Variability	
Mean	5.385027	Std Deviation	1.77453
Median	5.000000	Variance	3.14894
Mode	3.000000	Range	5.00000
		Interquartile Range	3.00000

The stress level data has a mean of 5.39 and a median of 5, indicating a relatively symmetric distribution. The standard deviation of 1.77 reflects moderate variability, while the interquartile range of 3 highlights the spread of the middle 50% of data. The skewness (0.15) and kurtosis (-1.33) suggest a slightly right-skewed and flatter-than-normal distribution. The range of 5 indicates some variation in stress levels among individuals.

```
/* 6. Heart Rate */
PROC UNIVARIATE DATA=HEALTH;
    VAR 'Heart Rate'n;
    HISTOGRAM / NORMAL;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBOX 'Heart Rate'n/ NAME='Boxplot For Heart Rate';
RUN;
```

Output:





The UNIVARIATE Procedure
Variable: Heart Rate

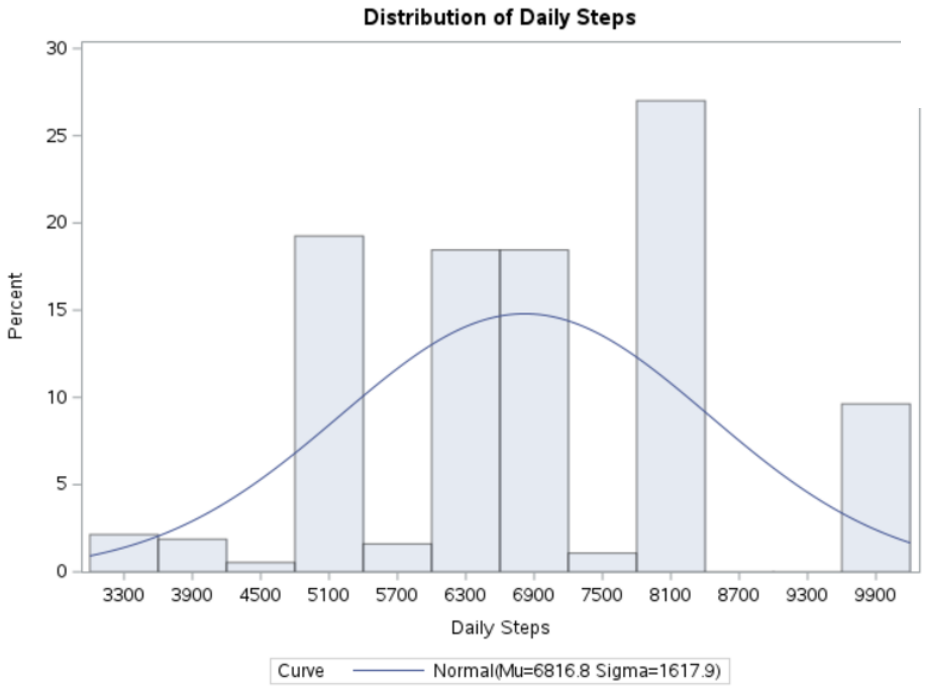
Moments			
N	374	Sum Weights	374
Mean	70.1657754	Sum Observations	26242
Std Deviation	4.13567554	Variance	17.1038121
Skewness	1.22482355	Kurtosis	2.28645467
Uncorrected SS	1847670	Corrected SS	6379.72193
Coeff Variation	5.89414927	Std Error Mean	0.21385069

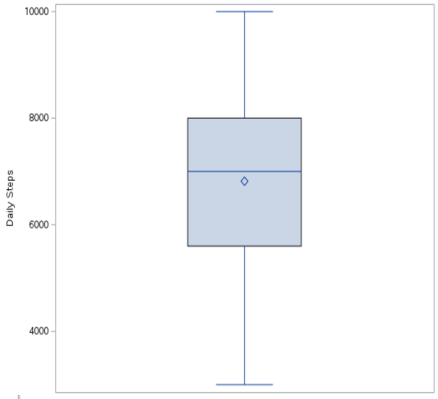
Basic Statistical Measures			
Location		Variability	
Mean	70.16578	Std Deviation	4.13568
Median	70.00000	Variance	17.10381
Mode	68.00000	Range	21.00000
		Interquartile Range	4.00000

The heart rate data shows a mean of 70.17 and a median of 70, indicating a relatively symmetric central tendency. However, the skewness (1.22) suggests a moderate right skew due to outliers above 80, as shown in the boxplot. The standard deviation of 4.14 indicates low variability, and the interquartile range (4) highlights a compact spread in the middle 50% of data. The histogram confirms a slight right-skewed distribution with most values clustering between 66 and 72.

```
/* 7. Daily Steps */
PROC UNIVARIATE DATA=HEALTH;
    VAR 'Daily Steps'n;
    HISTOGRAM / NORMAL;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBOX 'Daily Steps'n/ NAME='Boxplot For Daily Steps';
RUN;
```

Output:





The UNIVARIATE Procedure
Variable: Daily Steps

Moments			
N	374	Sum Weights	374
Mean	6816.84492	Sum Observations	2549500
Std Deviation	1617.91568	Variance	2617651.14
Skewness	0.17827733	Kurtosis	-0.3940306
Uncorrected SS	1.83559E10	Corrected SS	976383877
Coeff Variation	23.7340837	Std Error Mean	83.6604281

Basic Statistical Measures			
Location		Variability	
Mean	6816.845	Std Deviation	1618
Median	7000.000	Variance	2617651
Mode	8000.000	Range	7000
		Interquartile Range	2400

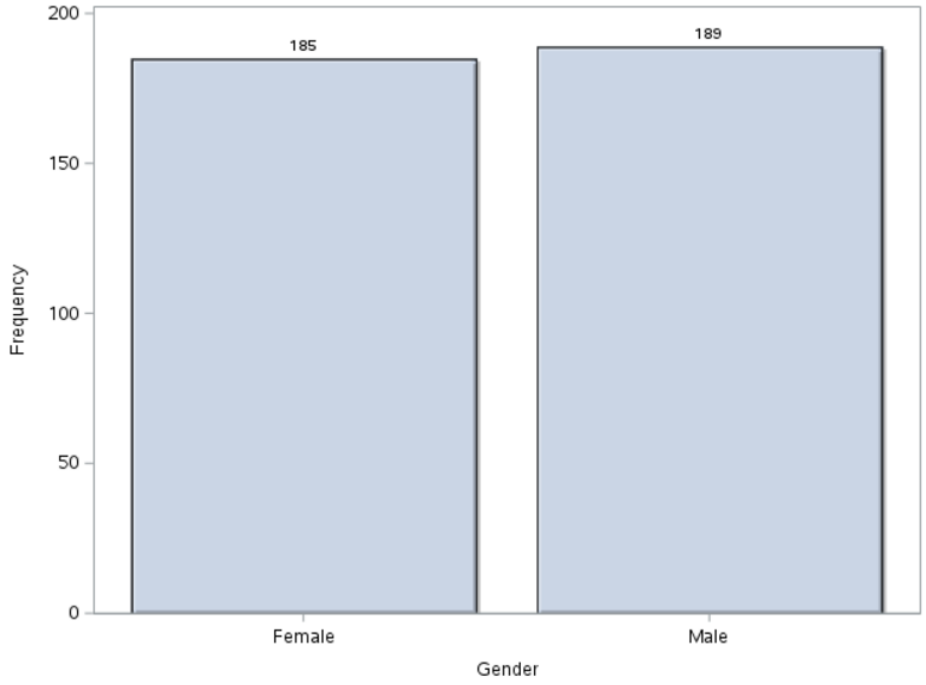
The daily steps data shows a mean of 6816.85 and a median of 7000, indicating a slightly left-skewed distribution with skewness of 0.18. The standard deviation of 1618 suggests moderate variability in daily steps. The interquartile range of 2400 highlights the spread of the middle 50% of the data. The histogram shows a nearly normal distribution with most values clustering between 5700 and 8100 steps, while the mode of 8000 steps suggests it is the most frequent value. The range of 7000 reflects substantial variability in step counts across the dataset.

```
/* Now let's perform for Categorical Variables */

/* 1. Gender */
PROC FREQ DATA=HEALTH;
    TABLES Gender;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBAR Gender / DATASKIN=CRISP DATALABEL;
RUN;
```

The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	185	49.47	185	49.47
Male	189	50.53	374	100.00

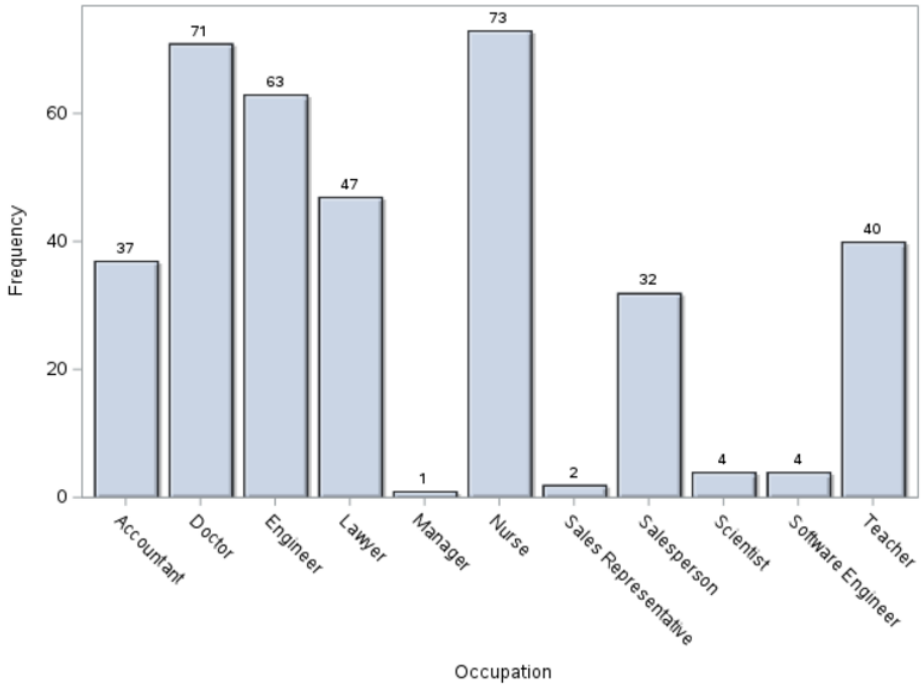
Output:



The gender distribution is nearly balanced, with 49.47% of participants identifying as female (185 individuals) and 50.53% identifying as male (189 individuals). The cumulative percentage indicates that both genders together make up 100% of the sample population. The bar chart reflects this proportional balance visually, with almost equal frequencies for males and females. This even distribution ensures gender representation in the data.

```sas
/* 2. Occupation */
PROC FREQ DATA=HEALTH;
    TABLES Occupation;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBAR Occupation / DATASKIN=CRISP DATALABEL;
RUN;
```

Output:



The FREQ Procedure

Occupation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Accountant	37	9.89	37	9.89
Doctor	71	18.98	108	28.88
Engineer	63	16.84	171	45.72
Lawyer	47	12.57	218	58.29
Manager	1	0.27	219	58.56
Nurse	73	19.52	292	78.07
Sales Representative	2	0.53	294	78.61
Salesperson	32	8.56	326	87.17
Scientist	4	1.07	330	88.24
Software Engineer	4	1.07	334	89.30
Teacher	40	10.70	374	100.00

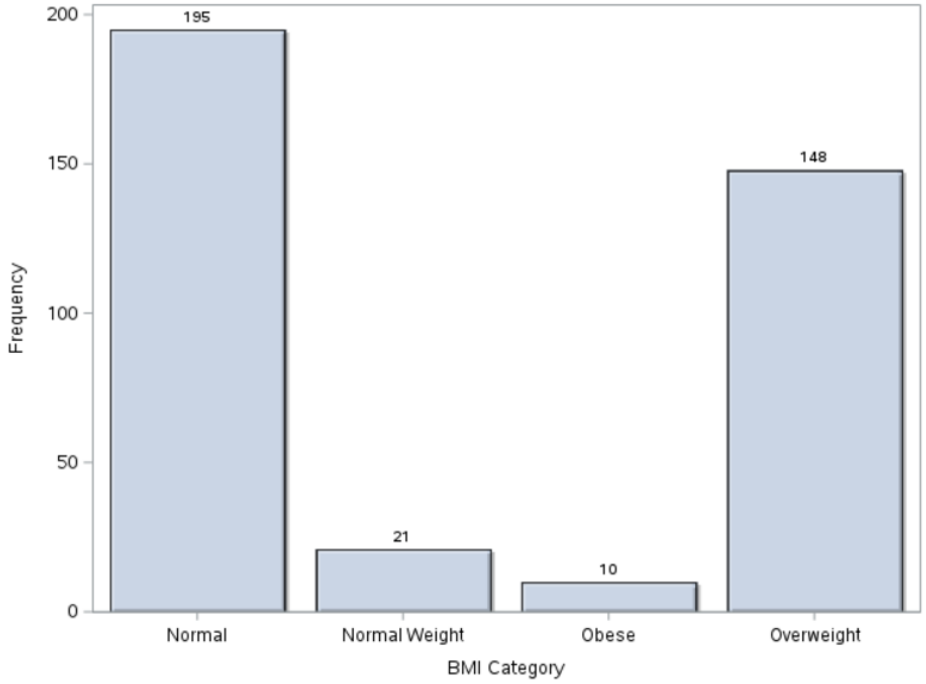
The occupation data indicates that the most common professions are nurses (73, 19.52%) and doctors (71, 18.98%), followed by engineers (63, 16.84%) and teachers (40, 10.7%). Lawyers (47, 12.57%) also represent a significant portion. Less common occupations include accountants (37, 9.89%) and salespeople (32, 8.56%), while managers, sales representatives, scientists, and software engineers have very low frequencies. The cumulative percentage shows that nurses, doctors, and engineers make up over half the sample population. This distribution highlights a strong representation in healthcare and engineering fields.

```
/* 3. BMI Category */
PROC FREQ DATA=HEALTH;
    TABLES 'BMI Category'n;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBAR 'BMI Category'n / DATASKIN=CRISP DATALABEL;
RUN;
```

Output:

The FREQ Procedure

BMI Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal	195	52.14	195	52.14
Normal Weight	21	5.61	216	57.75
Obese	10	2.67	226	60.43
Overweight	148	39.57	374	100.00



The BMI category distribution shows that the majority of individuals fall into the "Normal" category (52.14%, 195 individuals), followed by "Overweight" (39.57%, 148 individuals). A smaller proportion is categorized as "Normal Weight" (5.61%, 21 individuals), and only 2.67% (10 individuals) are classified as "Obese." The cumulative percentage indicates that nearly 60% of the sample has a BMI in the "Normal" or "Normal Weight" range, while overweight individuals account for a substantial portion. This distribution highlights the need to address weight management concerns for nearly 40% of the population.

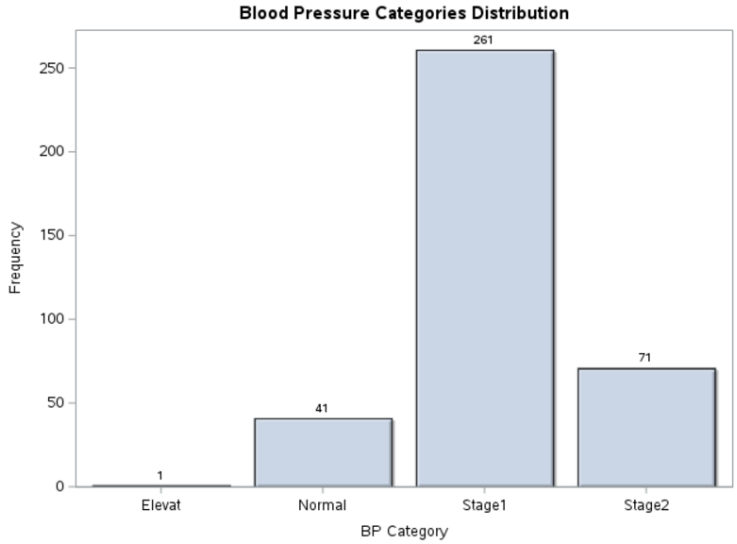
```
/* 4. Blood Pressure*/
/* Recategorizes Blood Pressure into bins */
data HEALTH;
   set HEALTH;

   /* Extract systolic and diastolic values from char type Blood Pressure */
   systolic = input(scan('Blood Pressure'n, 1, '/'), 8.);
   diastolic = input(scan('Blood Pressure'n, 2, '/'), 8.);

   /* Categorize Blood Pressure */
   if systolic < 120 and diastolic < 80 then 'BP Category'n = "Normal";
   else if systolic >= 120 and systolic <= 129 and diastolic < 80 then 'BP Category'n = "Elevated";
   else if (systolic >= 130 and systolic <= 139) or (diastolic >= 80 and diastolic <= 89) then 'BP Category'n =
   else if systolic >= 140 or diastolic >= 90 then 'BP Category'n = "Stage2";
   else if systolic > 180 or diastolic > 120 then 'BP Category'n = "Crisis";
   else 'BP Category'n = "Unknown"; /* In case of missing or incorrect values */
run;

proc freq data=HEALTH;
   tables 'BP Category'n;
   title "Distribution of Blood Pressure Categories";
run;

proc sgplot data=HEALTH;
   vbar 'BP Category'n / DATASKIN=CRISP DATALABEL;
   title "Blood Pressure Categories Distribution";
run;
```

Distribution of Blood Pressure Categories

The FREQ Procedure

BP Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Elevat	1	0.27	1	0.27
Normal	41	10.96	42	11.23
Stage1	261	69.79	303	81.02
Stage2	71	18.98	374	100.00

Output:



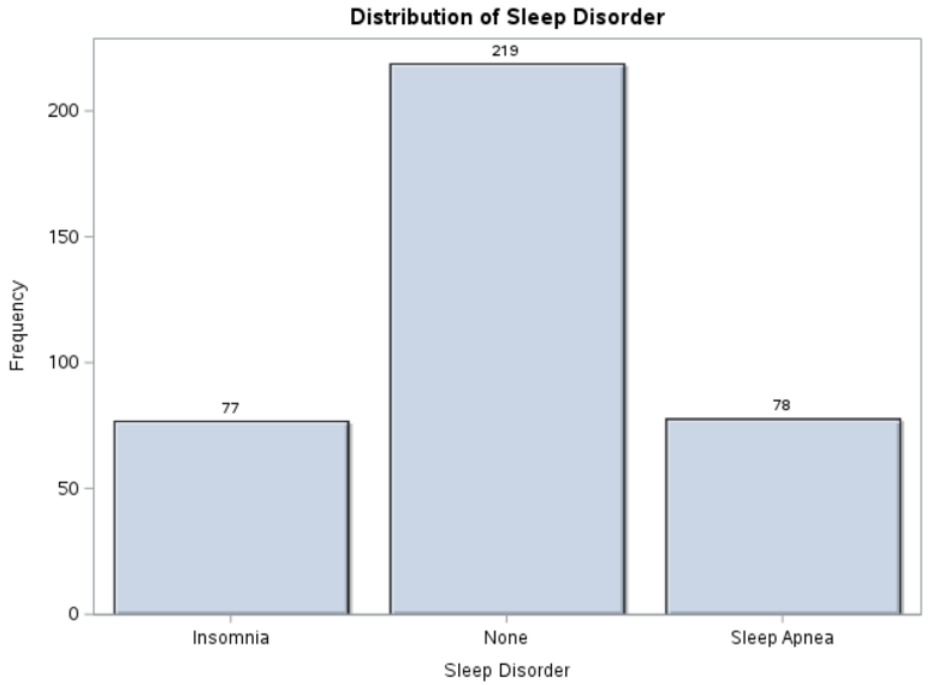
The blood pressure data was preprocessed to convert numerical values into categorical labels for better analysis, resulting in four categories: "Elevated," "Normal," "Stage 1," and "Stage 2." The majority of individuals fall into the "Stage 1" category (69.79%, 261 individuals), followed by "Stage 2" (18.98%, 71 individuals) and "Normal" (10.96%, 41 individuals). Only one individual (0.27%) was classified as "Elevated." This categorization highlights that most participants have elevated blood pressure requiring attention, with a smaller proportion maintaining normal levels.

```sas
/* 5. Sleep Disorder */
PROC FREQ DATA=HEALTH;
    TABLES 'Sleep Disorder'n;
RUN;

PROC SGPLOT DATA=HEALTH;
    VBAR 'Sleep Disorder'n / DATASKIN=CRISP DATALABEL;
    title "Distribution of Sleep Disorder";
RUN;
```

Output:

The FREQ Procedure

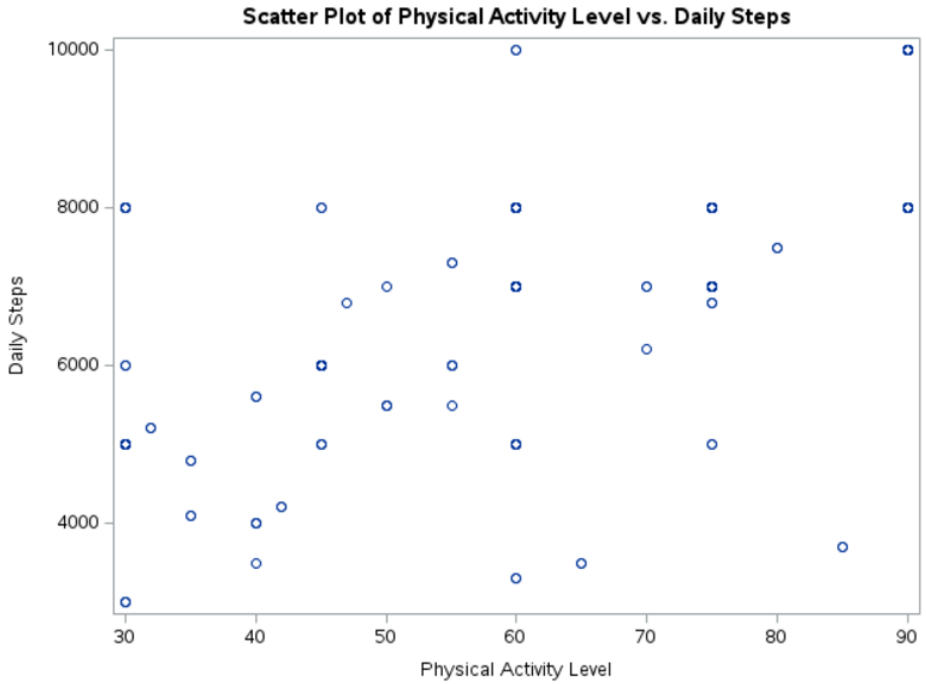
Sleep Disorder	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Insomnia	77	20.59	77	20.59
None	219	58.56	296	79.14
Sleep Apnea	78	20.86	374	100.00



The data reveals that 58.56% of individuals (219 people) do not report any sleep disorders, while 20.59% (77 individuals) suffer from insomnia and 20.86% (78 individuals) experience sleep apnea. This shows that while the majority have healthy sleep patterns, a significant proportion (41.45%) face sleep-related issues. These findings highlight the need for focused interventions to address and manage insomnia and sleep apnea effectively.

```
/* Now let's plot a relationship between two variables */
PROC SGPLOT DATA=HEALTH;
    /* Scatter plot for Physical Activity Level vs. Daily Steps */
    SCATTER X='Physical Activity Level'n Y='Daily Steps'n;
    TITLE "Scatter Plot of Physical Activity Level vs. Daily Steps";
RUN;
```

Output:



The scatter plot shows the relationship between physical activity level and daily steps. There appears to be a positive trend, as higher physical activity levels generally correspond to a greater number of daily steps. However, the data points are widely scattered, indicating variability in the relationship. While some individuals with lower physical activity levels still achieve high daily steps, the overall pattern suggests that increased physical activity levels tend to align with a higher step count. This variability might reflect differing definitions or measurements of physical activity beyond step counts.

```
/* Creating a Joint Distribution, to get more information about the variable relation */
PROC FREQ DATA=HEALTH;
    TABLES Gender * 'Sleep Duration'n ;
    Title "joint Distribution of Gender and Sleep Duration";
RUN;

PROC FREQ DATA=HEALTH;
    TABLES Gender * Occupation ;
    Title "joint Distribution of Gender and Occupation";
RUN;

PROC FREQ DATA=HEALTH;
    TABLES Gender * 'BMI Category'n ;
    Title "joint Distribution of Gender and BMI Category";
RUN;
```

Output:

joint Distribution of Gender and Occupation

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Gender by Occupation

Gender	Occupation											
	Accountant	Doctor	Engineer	Lawyer	Manager	Nurse	Sales Representative	Salesperson	Scientist	Software Engineer	Teacher	Total
Female	36 9.63 19.46 97.30	2 0.53 1.08 2.82	32 8.56 17.30 50.79	2 0.53 1.08 4.26	1 0.27 0.54 100.00	73 19.52 39.46 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 1.07 2.16 100.00	0 0.00 0.00 0.00	35 9.36 18.92 87.50	185 49.47
Male	1 0.27 0.53 2.70	69 18.45 36.51 97.18	31 8.29 16.40 49.21	45 12.03 23.81 95.74	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.53 1.06 100.00	32 8.56 16.93 100.00	0 0.00 0.00 0.00	4 1.07 2.12 100.00	5 1.34 2.65 12.50	189 50.53
Total	37 9.89	71 18.98	63 16.84	47 12.57	1 0.27	73 19.52	2 0.53	32 8.56	4 1.07	4 1.07	40 10.70	374 100.00

The joint distribution of gender and occupation shows that females dominate roles like nursing (73 out of 73) and teaching (87.5% of teachers), while males are more represented as doctors (97.18%) and lawyers (95.74%). Engineers and salespersons show a relatively balanced gender representation, highlighting variations in gender distribution across professions.

joint Distribution of Gender and BMI Category

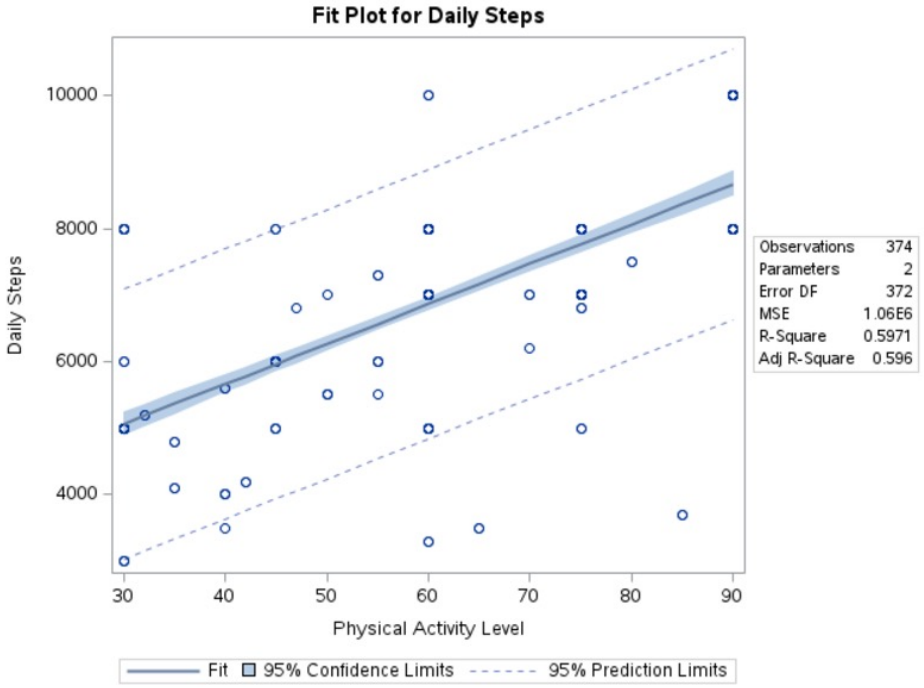
The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Gender by BMI Category				
	BMI Category				
Gender	Normal	Normal Weight	Obese	Overweight	Total
Female	64 17.11 34.59 32.82	14 3.74 7.57 66.67	1 0.27 0.54 10.00	106 28.34 57.30 71.62	185 49.47
Male	131 35.03 69.31 67.18	7 1.87 3.70 33.33	9 2.41 4.76 90.00	42 11.23 22.22 28.38	189 50.53
Total	195 52.14	21 5.61	10 2.67	148 39.57	374 100.00

The joint distribution of gender and BMI category shows that the majority of males (69.31%) and females (57.30%) fall into the "Normal" and "Overweight" categories combined. Females have a slightly higher representation in the "Overweight" category (71.62%), while males dominate the "Normal" category (67.18%). The "Obese" category has a low prevalence for both genders, with males (2.41%) slightly exceeding females (0.27%). This distribution highlights gender differences in BMI classifications, with overweight being more common in females and normal weight more common in males.

```
/* Calculate the least squares regression line for y in terms of x. How reliable is
this model, justify? Provide an equation. Plot the regression line. */

/* x = Physical Activity Level */
/* y = Daily Steps */
/* Perform regression analysis */
PROC REG DATA=HEALTH;
    MODEL 'Daily Steps'n = 'Physical Activity Level'n;
    /* Output predicted values to a new dataset for plotting */
    OUTPUT OUT=predicted_data P=Predicted_Steps;
RUN;
```

Output:



The fit plot shows a positive linear relationship between physical activity level and daily steps, with an R-squared value of 0.5971 indicating that about 59.7% of the variation in daily steps is explained by physical activity level. The 95% confidence and prediction intervals highlight variability around the fit line. While higher activity levels generally correspond to more steps, the scatter suggests other factors also influence step counts.

```
/* Correlation of other variables with physical activity */
proc corr data=HEALTH pearson;
   var 'Physical Activity Level'n;
   with 'Age'n 'Sleep Duration'n 'Quality of Sleep'n 'Stress Level'n 'Heart Rate'n 'Daily Steps'n ;
   title "Correlation of Variables with Physical Activity Level";
run;
```

Output:

Pearson Correlation Coefficients, N = 374 Prob > |r| under H0: Rho=0	
	Physical Activity Level
Age	0.17899 0.0005
Sleep Duration	0.21236 <.0001
Quality of Sleep	0.19290 0.0002
Stress Level	-0.03413 0.5105
Heart Rate	0.13697 0.0080
Daily Steps	0.77272 <.0001

The correlation table shows the strength and significance of the relationship between physical activity level and various factors.

- A strong positive correlation is observed between physical activity level and daily steps ($r = 0.77272$, $p < 0.0001$), indicating that increased physical activity is strongly associated with higher daily steps.
- Moderate positive correlations are seen with sleep duration ($r = 0.21236$, $p < 0.0001$) and quality of sleep ($r = 0.19290$, $p = 0.0002$), suggesting better sleep patterns are associated with higher activity levels.
- Weak correlations are observed with age ($r = 0.17899$, $p = 0.0005$) and heart rate ($r = 0.13697$, $p = 0.0080$), while stress level shows a negligible and insignificant negative correlation ($r = -0.03413$, $p = 0.5105$).

In conclusion, physical activity level is most strongly related to daily steps, with moderate associations to sleep quality and duration, and negligible influence from stress levels.

```
/* Correlation of other variables with Sleep Duration */
proc corr data=HEALTH pearson;
   var 'Sleep Duration'n;
   with 'Age'n 'Quality of Sleep'n 'Physical Activity Level'n 'Stress Level'n 'Heart Rate'n 'Daily Steps'n ;
   title "Correlation of Variables with Sleep Duration";
run;
```

Output:

Pearson Correlation Coefficients, N = 374 Prob > |r| under H0: Rho=0	Sleep Duration
Age	0.34471 <.0001
Quality of Sleep	0.88321 <.0001
Physical Activity Level	0.21236 <.0001
Stress Level	-0.81102 <.0001
Heart Rate	-0.51645 <.0001
Daily Steps	-0.03953 0.4459

The correlation table shows the relationship between sleep duration and various factors.

- Sleep duration has a strong positive correlation with the quality of sleep ($r = 0.88321$, $p < 0.0001$), indicating that longer sleep is closely associated with better sleep quality.
- A moderate positive correlation exists with age ($r = 0.34471$, $p < 0.0001$) and a weak positive correlation with physical activity level ($r = 0.21236$, $p < 0.0001$).
- However, sleep duration is negatively correlated with stress level ($r = -0.81102$, $p < 0.0001$) and heart rate ($r = -0.51645$, $p < 0.0001$), suggesting that longer sleep is linked to reduced stress and lower heart rate.
- The correlation with daily steps is negligible and insignificant ($r = -0.03953$, $p = 0.4459$).

In conclusion, sleep duration is most strongly linked to sleep quality and stress level, highlighting the importance of stress management for better sleep. Other factors, like physical activity and heart rate, show weaker associations.

```
/* Correlation of other variables with Quality of Sleep */
proc corr data=HEALTH pearson;
   var 'Quality of Sleep'n;
   with 'Age'n 'Sleep Duration'n 'Physical Activity Level'n 'Stress Level'n 'Heart Rate'n 'Daily Steps'n ;
   title "Correlation of Variables with Quality of Sleep";
run;
```

Output:

Pearson Correlation Coefficients, N = 374 Prob > |r| under H0: Rho=0	
	Quality of Sleep
Age	0.47373 <.0001
Sleep Duration	0.88321 <.0001
Physical Activity Level	0.19290 0.0002
Stress Level	-0.89875 <.0001
Heart Rate	-0.65986 <.0001
Daily Steps	0.01679 0.7462

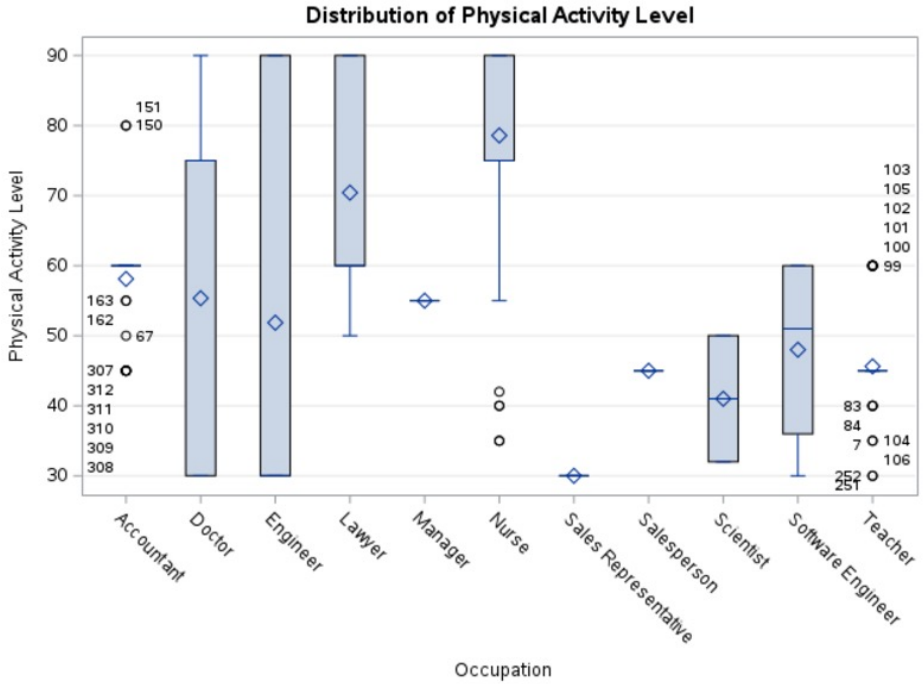
The correlation table reveals the relationships between sleep quality and several factors.

- Sleep quality is strongly and positively correlated with sleep duration ($r = 0.88321$, $p < 0.0001$), indicating that better sleep quality is closely associated with longer sleep duration.
- A moderate positive correlation is observed with age ($r = 0.47373$, $p < 0.0001$), while a weak positive correlation exists with physical activity level ($r = 0.19290$, $p = 0.0002$).
- Stress level shows a strong negative correlation with sleep quality ($r = -0.89875$, $p < 0.0001$), suggesting that higher stress significantly reduces sleep quality.
- Similarly, heart rate has a moderately negative correlation ($r = -0.65986$, $p < 0.0001$), while the correlation with daily steps is negligible and not significant ($r = 0.01679$, $p = 0.7462$).

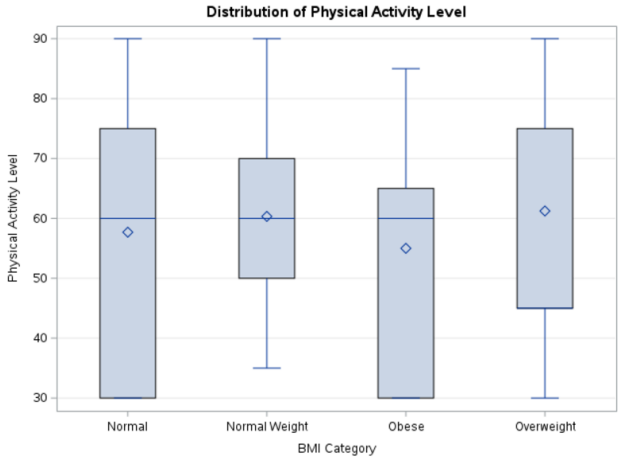
In conclusion, sleep quality is most strongly influenced by sleep duration and stress level, emphasizing the importance of managing stress and ensuring sufficient sleep for better sleep quality. Other factors like physical activity and heart rate show moderate to weak associations.

```
/* ANOVA to analyze the effect of categorical variables on physical activity */
proc glm data=HEALTH;
   class Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n;
   model 'Physical Activity Level'n = Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n;
   means Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n / tukey;
   title "Effect of Categorical Variables on Physical Activity Level";
run;
quit;
```

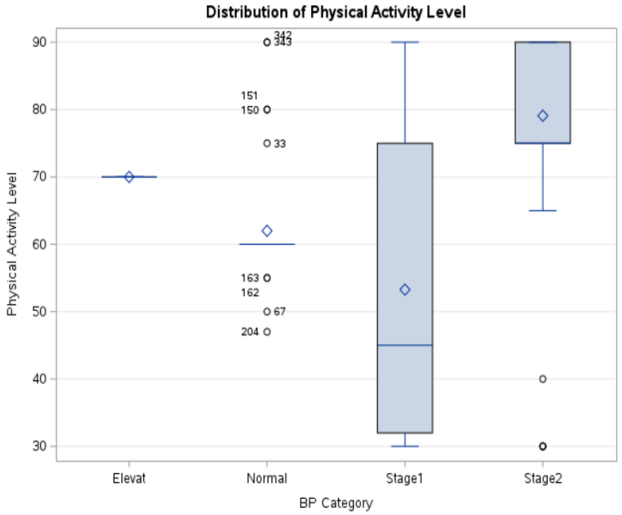
Output:



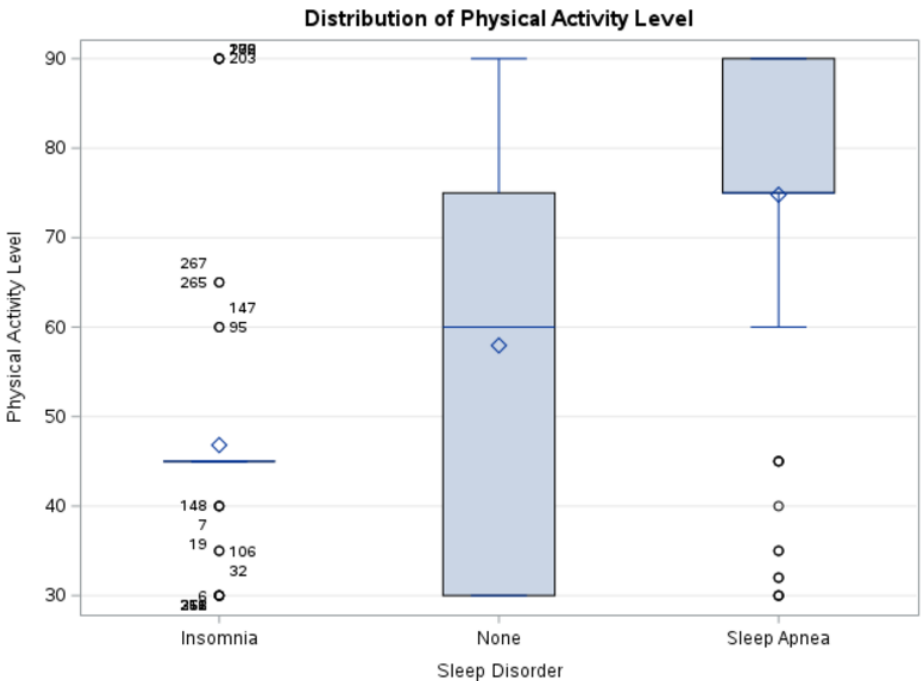
The boxplot displays the distribution of physical activity levels across various occupations. Engineers, doctors, and teachers exhibit wider variability in physical activity levels, as seen from their larger interquartile ranges. Nurses and sales representatives show relatively consistent physical activity levels with narrow distributions. Outliers are present in several occupations, such as doctors, engineers, and software engineers, indicating individuals with unusually high or low activity levels. Overall, physical activity levels vary significantly by occupation, with some roles showing higher consistency than others.

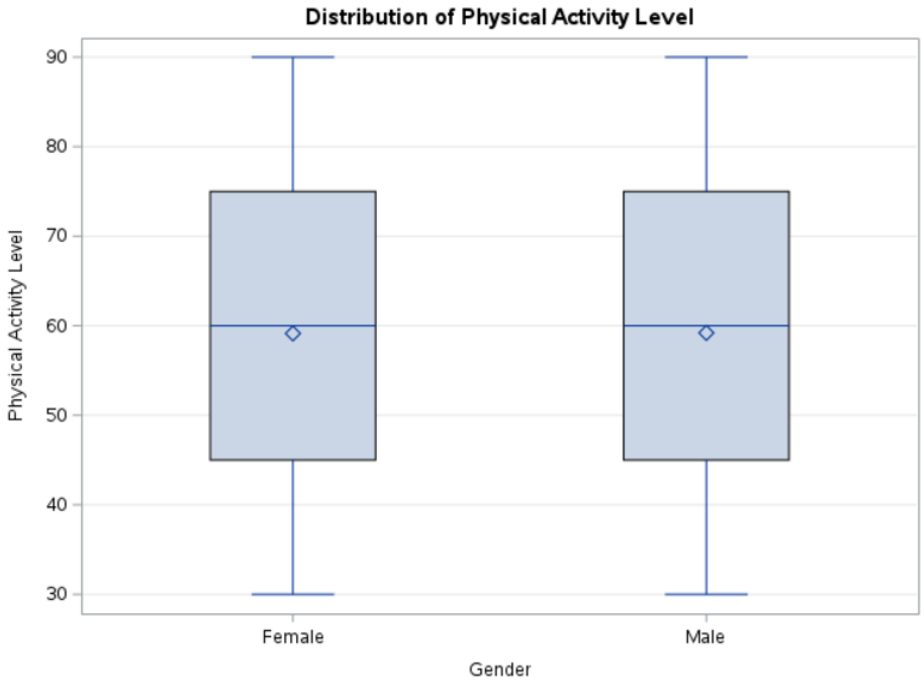


The boxplot illustrates the distribution of physical activity levels across BMI categories. Individuals in the "Normal" BMI category exhibit the widest range of activity levels, while "Normal Weight" and "Obese" categories show narrower distributions. The median physical activity level is similar across categories, but the variability in the "Normal" and "Overweight" groups suggests greater diversity in activity levels within these categories. Outliers are present in most groups, highlighting individuals with unusually high or low physical activity.



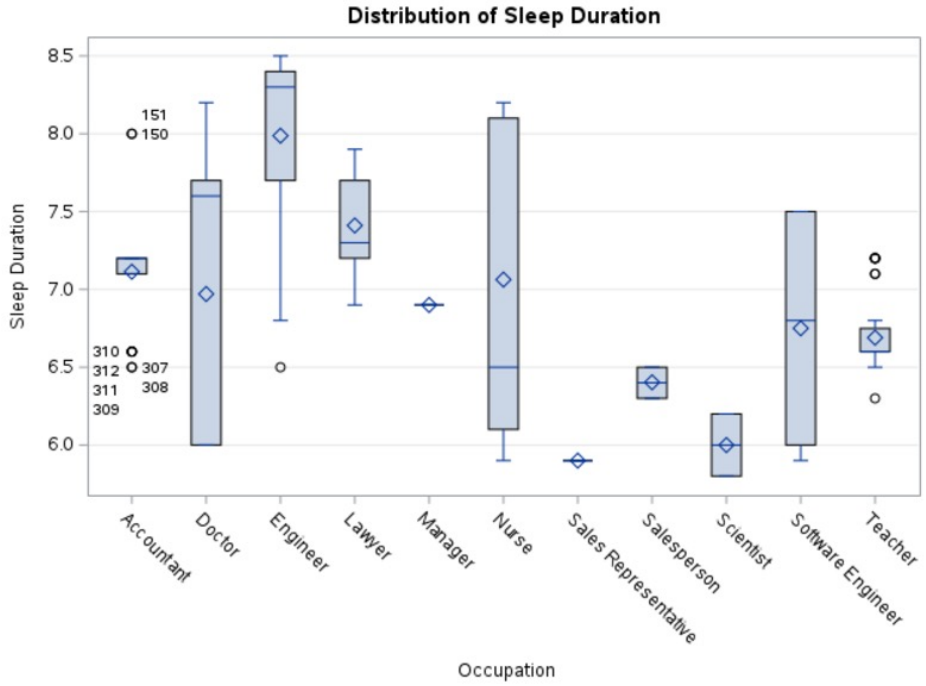
The boxplot shows the distribution of physical activity levels across different blood pressure (BP) categories. "Stage 1" has the widest range and variability in activity levels, while "Stage 2" shows higher median activity with less variability. The "Normal" and "Elevated" categories display narrow distributions with relatively consistent activity levels. Outliers are present in all categories, especially in "Stage 1," indicating individuals with unusually high or low physical activity.


Distribution of Physical Activity Level
Physical Activity Level
90
80
70
60
50
40
30
267
265
147
95
148
7
19
106
32
Insomnia
None
Sleep Apnea
Sleep Disorder

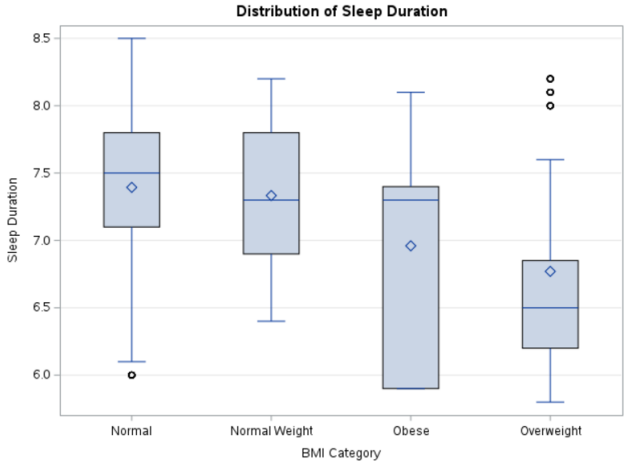

Distribution of Physical Activity Level
Physical Activity Level
90
80
70
60
50
40
30
Female
Male
Gender

```
/* ANOVA to analyze the effect of categorical variables on Sleep Duration */
proc glm data=HEALTH;
   class Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n;
   model 'Sleep Duration'n = Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n;
   means Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n / tukey;
   title "Effect of Categorical Variables on Sleep Duration";
run;
quit;
```

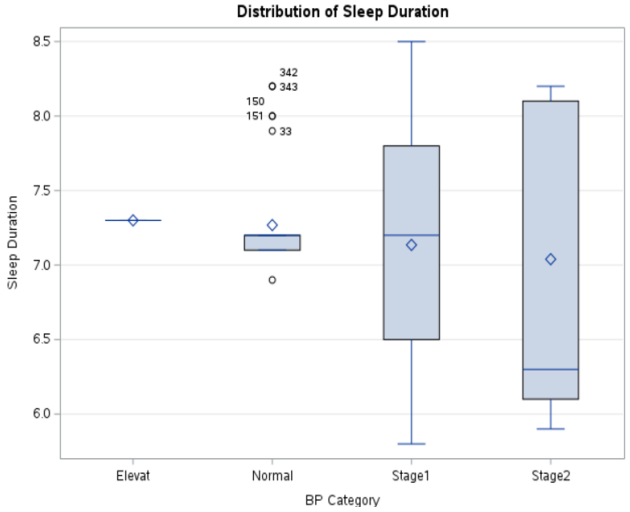
Output:



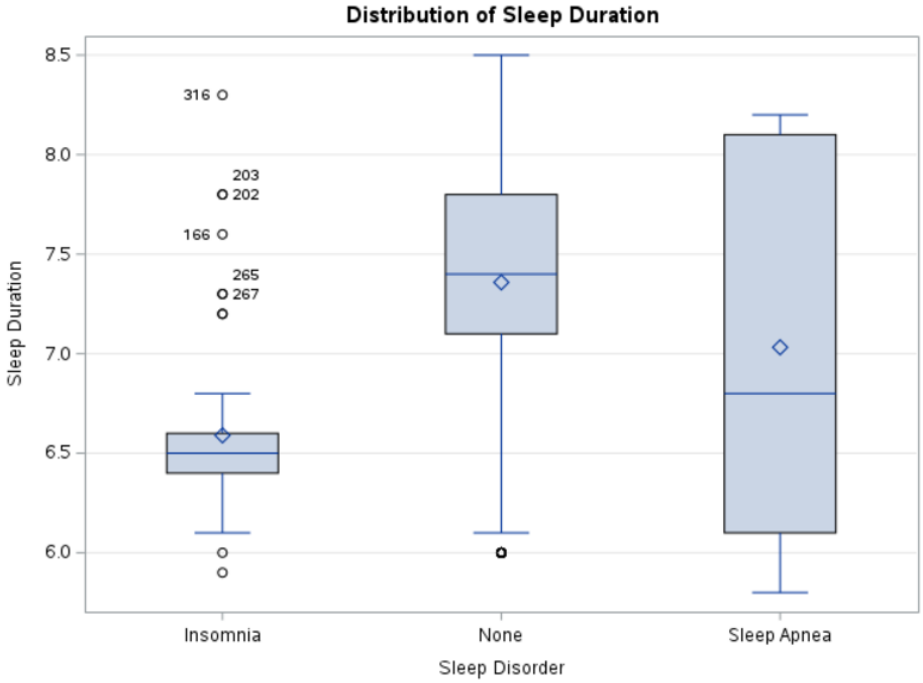
The boxplot displays the distribution of sleep duration across various occupations. Engineers and nurses exhibit wider variability in sleep duration, as indicated by their larger interquartile ranges. Nurses have the highest median sleep duration, while sales representatives and scientists show lower median sleep durations with narrow variability. Occupations like teachers and software engineers display moderate variability but include outliers with unusually low or high sleep durations. Overall, the plot highlights significant differences in sleep patterns across occupations, likely influenced by job-specific demands and schedules.

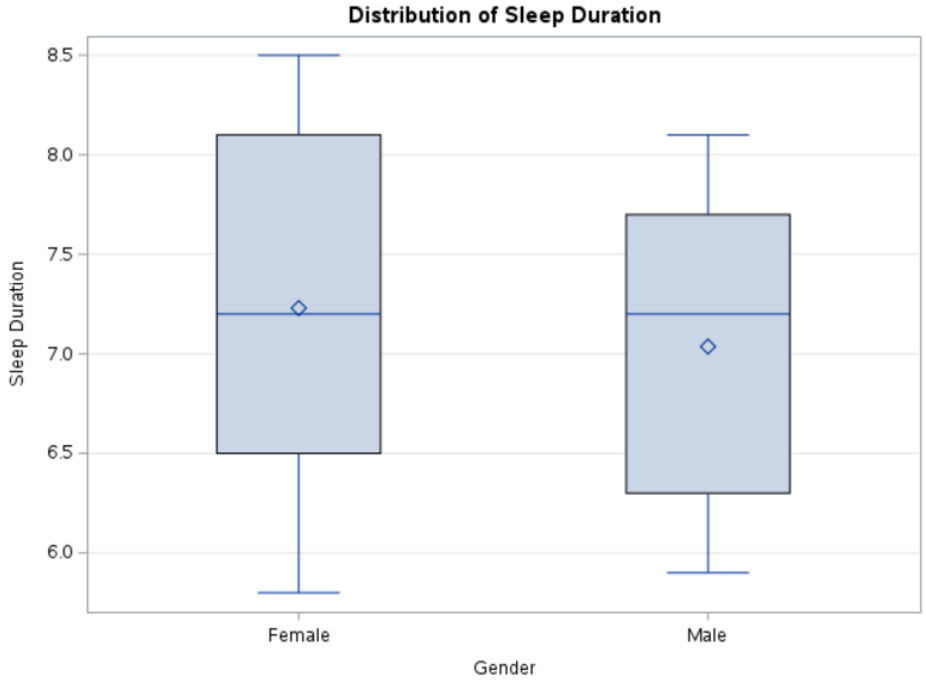


The boxplot shows the distribution of sleep duration across BMI categories. Individuals in the "Normal" and "Normal Weight" BMI categories have relatively consistent sleep durations with median values around 7.5 hours. The "Obese" category exhibits the widest variability, with sleep durations ranging from 6 to over 8 hours. The "Overweight" category shows shorter median sleep durations and several outliers with unusually high values. This distribution highlights that sleep duration tends to vary more in individuals with higher BMI levels, especially in the "Obese" category.



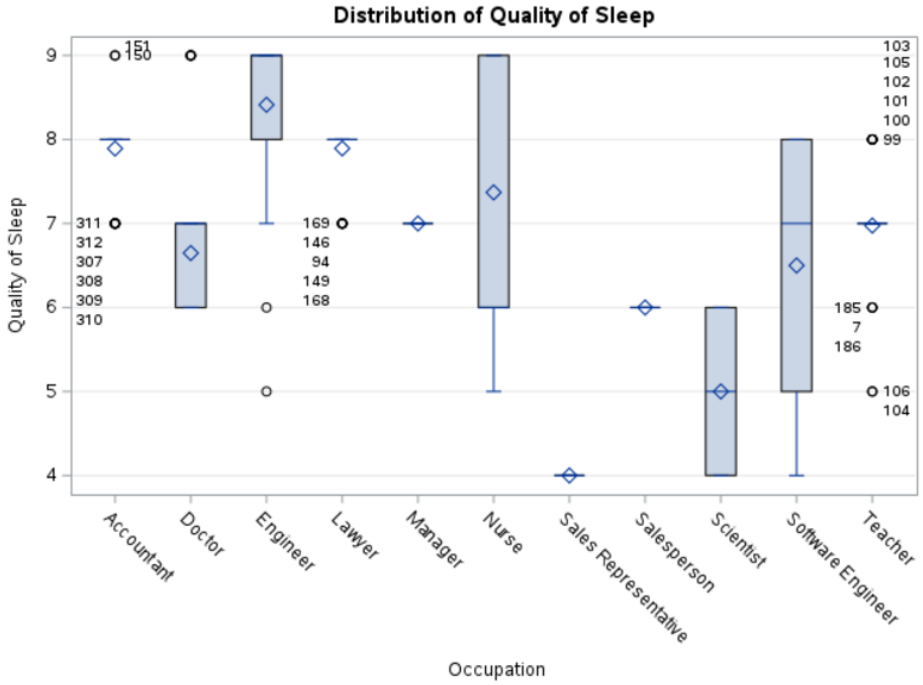
The boxplot illustrates the distribution of sleep duration across blood pressure (BP) categories. The "Stage 2" category shows the widest variability in sleep duration, with a lower median and a broader range compared to other categories. The "Stage 1" category also exhibits significant variability, while "Normal" and "Elevated" categories have narrower distributions and relatively consistent sleep durations. Outliers are visible in the "Normal" category, representing individuals with unusually long or short sleep durations. Overall, sleep duration appears to vary more in individuals with elevated BP stages.


Distribution of Sleep Duration
Sleep Duration
8.5
8.0
7.5
7.0
6.5
6.0
316
203
202
166
265
267
Insomnia
None
Sleep Apnea
Sleep Disorder

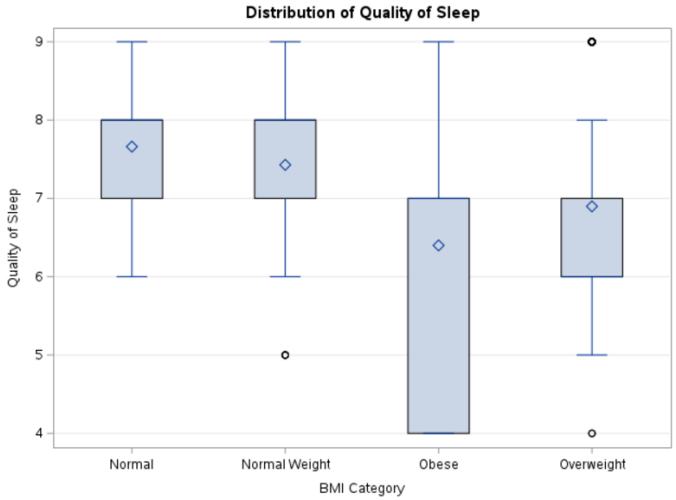

Distribution of Sleep Duration
Sleep Duration
8.5
8.0
7.5
7.0
6.5
6.0
Female
Male
Gender

```
/* ANOVA to analyze the effect of categorical variables on Quality of Sleep */
proc glm data=HEALTH;
   class Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n;
   model 'Quality of Sleep'n = Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n;
   means Gender Occupation 'BMI Category'n 'BP Category'n 'Sleep Disorder'n / tukey;
   title "Effect of Categorical Variables on Quality of Sleep";
run;
quit;
```

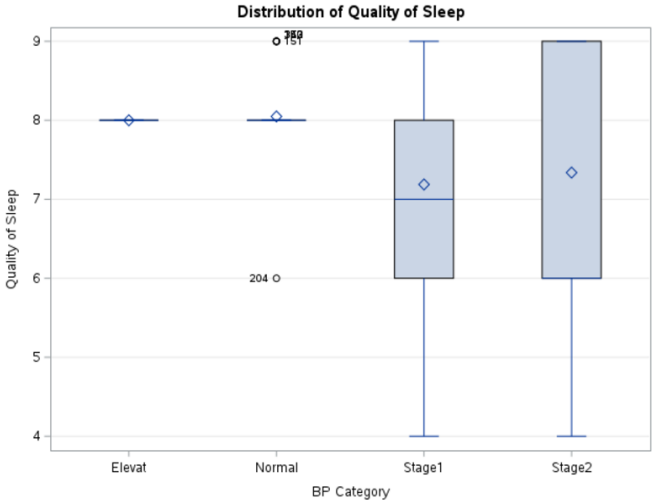
Output:



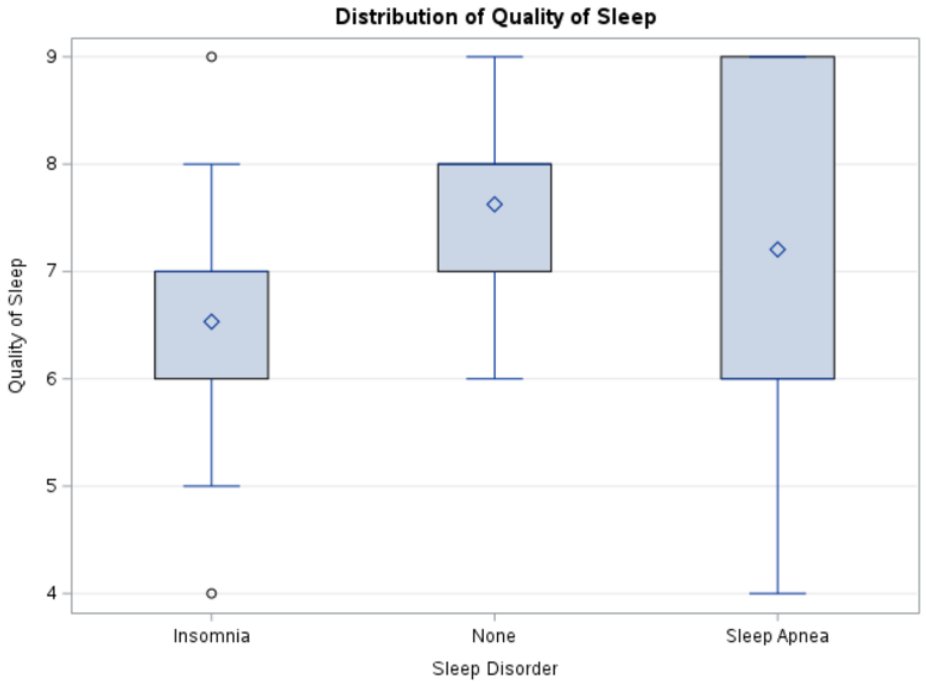
The boxplot displays the distribution of sleep quality across different occupations. Nurses show the widest range of sleep quality, with a median slightly higher than other occupations, while engineers and teachers exhibit more moderate variability. Sales representatives and scientists have narrower distributions, indicating consistent sleep quality within these roles. Outliers are evident in several occupations, including doctors, software engineers, and teachers, representing individuals with unusually high or low sleep quality. Overall, sleep quality varies significantly among occupations, influenced likely by job demands and schedules.

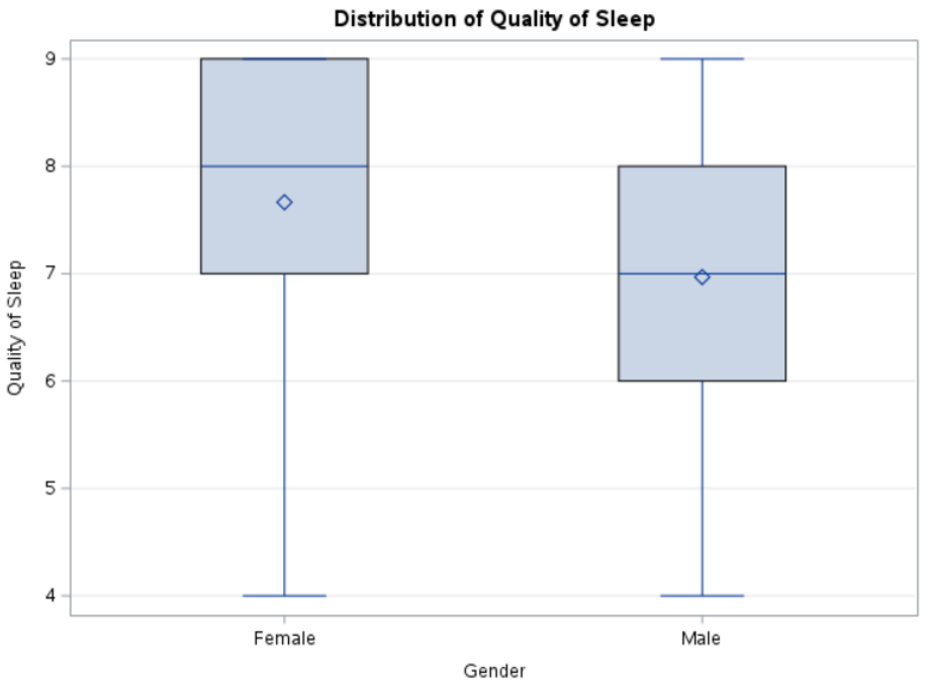


The boxplot illustrates the distribution of sleep quality across BMI categories. Individuals with "Obese" BMI exhibit the widest variability in sleep quality, with lower median values compared to other categories. The "Normal" and "Normal Weight" categories show similar distributions with consistent sleep quality and higher median values. The "Overweight" category has moderate variability and a slightly lower median sleep quality. Outliers are present in all groups, especially in the "Obese" and "Overweight" categories, indicating individuals with exceptionally high or low sleep quality. This highlights a potential association between BMI and sleep quality.



The boxplot depicts the distribution of sleep quality across different blood pressure (BP) categories. The "Stage 2" category exhibits the widest variability in sleep quality, with a lower median compared to other categories. The "Stage 1" category also shows considerable variability, whereas the "Elevated" and "Normal" BP categories have narrow distributions with higher and consistent medians. Outliers are present, particularly in the "Normal" and "Stage 2" categories, representing individuals with notably different sleep quality. This suggests that higher BP stages may be associated with greater variability and potential declines in sleep quality.


Distribution of Quality of Sleep
Quality of Sleep
9
8
7
6
5
4
Insomnia
None
Sleep Apnea
Sleep Disorder


Distribution of Quality of Sleep
Quality of Sleep
9
8
7
6
5
4
Female
Male
Gender

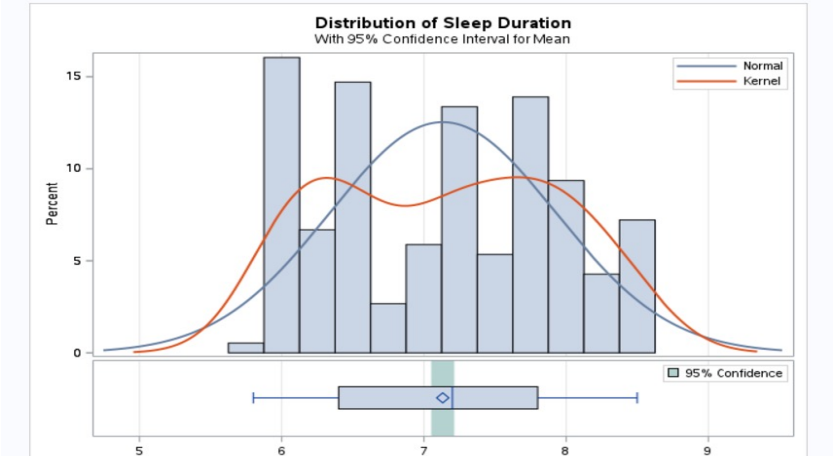
```
/* Now let's perform different types of One Sample test on our Dataset */

/* 1. One Sample t-test */

/* The one-sample t-test is used to compare the mean of a continuous variable against a hypothesized value.
In this case, we will test if the average Sleep Duration is equal to the recommended 7 hours.

Hypotheses
Null Hypothesis (H₀): The mean sleep duration is 7 hours μ=7
Alternative Hypothesis (H₁): The mean sleep duration is not 7 hours μ≠7.*/

PROC TTEST DATA=HEALTH H0=7 ALPHA=0.05;
    VAR 'Sleep Duration'n;
RUN;
```

Output:

N	Mean	Std Dev	Std Err	Minimum	Maximum
374	7.1321	0.7957	0.0411	5.8000	8.5000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
7.1321	7.0512	7.2130	0.7957	0.7424	0.8572

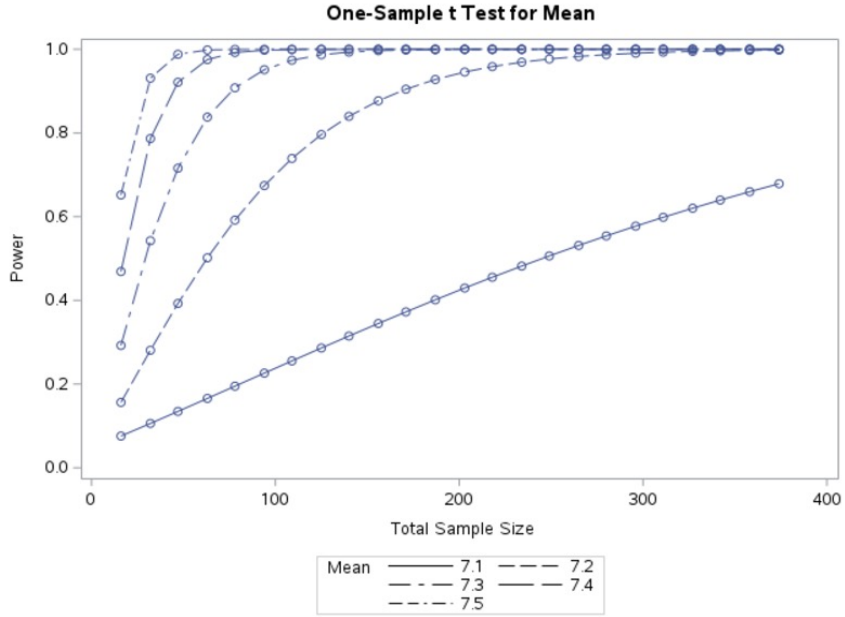
DF	t Value	Pr > |t|
373	3.21	0.0014



The histogram and boxplot illustrate the distribution of sleep duration with key statistics. The mean sleep duration is 7.13 hours, with a standard deviation of 0.80 hours, indicating moderate variability. The 95% confidence interval for the mean ranges from 7.05 to 7.21 hours, showing statistical reliability. The t-test result (t = 3.21, p = 0.0014) suggests that the mean is significantly different from a hypothesized value. The normal curve aligns with the histogram, while the kernel density highlights subtle deviations. The boxplot confirms no extreme outliers, providing a comprehensive view of the data's spread and central tendency. From the t-test results (t = 3.21, p = 0.0014), the p-value is significantly less than 0.05, leading us to reject the null hypothesis. Therefore, we conclude that the average sleep duration in the dataset is statistically different from 7 hours.

```
/* Power Calculation for Multiple Alternative Means in One-Sample t-Test */
ODS GRAPHICS ON;
PROC POWER;
    ONESAMPLEMEANS TEST=T
        MEAN = 7.1 7.2 7.3 7.4 7.5 /* List of alternative means */
        STDDEV = 0.7957 /* Standard Deviation from the dataset */
        NTOTAL = 374 /* Total sample size */
        ALPHA = 0.05 /* Significance level */
        NULLMEAN = 7 /* Hypothesized Mean */
        SIDES=2
        POWER=.;
    PLOT X=N MIN=1 MAX=374 NPOINTS=25;
RUN;
ODS GRAPHICS OFF;
```

Output:



Computed Power		
Index	Mean	Power
1	7.1	0.679
2	7.2	0.998
3	7.3	>.999
4	7.4	>.999
5	7.5	>.999

The power analysis for the one-sample t-test indicates how likely the test is to detect a true difference from the null hypothesis (H0:μ=7) at various means. As the sample size increases, the power also increases significantly for all means, with higher mean values (e.g., 7.3, 7.4, 7.5) reaching a power greater than 0.999 even for relatively smaller sample sizes. This analysis shows that the test is highly sensitive to deviations from the null hypothesis, especially with larger samples or larger effect sizes.