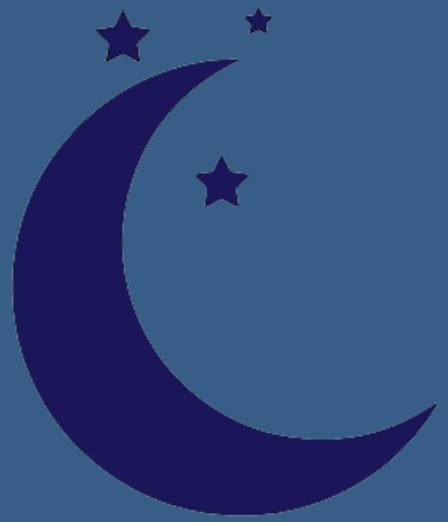


PREDICTING SLEEP DURATION

Ava Nafisi
904-514-1798
11/15/2023

PROBLEM INTRODUCTION

- According to the National Institutes of Health, adults within the United States on average sleep less than seven hours each night.
- This is less than the expert recommended amount of seven to nine hours.
- While the number of hours a person sleeps is usually by individual choice, there may be unexpected factors or trends that relate to the number of hours a person is getting each night.

ABOUT THE DATA

TITLE

The dataset used for this project is the Sleep and Health and Lifestyle Dataset

SOURCE

The dataset was located on Kaggle and created by Laksika Tharmalingam

SIZE

The data contains 374 observations of different individuals.

There are 13 columns, with one column containing the unique id's of each individual.

ENVIRONMENT

All statistical analysis in the presentation was done using R through R Studio v1.4

DATASET

NUMERICAL (6)

- **Age** - age of person (years)
- **Quality.of.Sleep** - subjective numerical rating of their quality of sleep (1-10)
- **Physical.Activity.Level** - number of daily physical activity (minutes/day)
- **Stress.Level** - subjective numerical rating of stress level (1-10)
- **Heart.Rate** - resting heart rate (bpm)
- **Daily.Steps** - count of steps taken per day.

CATEGORICAL (5)

- **Gender** - Gender of person (Male/Female)
- **Occupation** - name of person's occupation (10 Categories)
- **BMI.Category** - BMI category of person (Normal, Overweight, Obese)
- **Blood.Pressure** – individual's blood pressure (1 – high, 0 – normal)
- **Sleep.Disorder**: If the person has a sleep disorder (None, Insomnia, Sleep Apnea)

RESPONSE

Sleep.Duration – the number of hours a person sleeps per night (hours)

DATASET

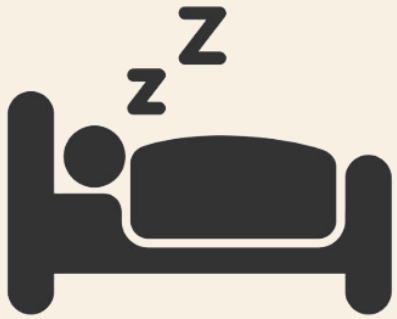
	Person.ID <int>	Gender <fctr>	Age <int>	Occupation <fctr>	Sleep.Duration <dbl>	Quality.of.Sleep <int>	Physical.Activity.Level <int>
1	1	Male	27	Software Engineer	6.1	6	42
2	2	Male	28	Doctor	6.2	6	60
3	3	Male	28	Doctor	6.2	6	60
4	4	Male	28	Sales Representative	5.9	4	30
5	5	Male	28	Sales Representative	5.9	4	30
6	6	Male	28	Software Engineer	5.9	4	30
7	7	Male	29	Teacher	6.3	6	40
8	8	Male	29	Doctor	7.8	7	75
9	9	Male	29	Doctor	7.8	7	75
10	10	Male	29	Doctor	7.8	7	75

Stress.Level <int>	BMI.Category <fctr>	Blood.Pressure <int>	Heart.Rate <int>	Daily.Steps <int>	Sleep.Disorder <fctr>
6	Overweight	0	77	4200	None
8	Normal	0	75	10000	None
8	Normal	0	75	10000	None
8	Obese	1	85	3000	Sleep Apnea
8	Obese	1	85	3000	Sleep Apnea
8	Obese	1	85	3000	Insomnia
7	Obese	1	82	3500	Insomnia
6	Normal	0	70	8000	None
6	Normal	0	70	8000	None
6	Normal	0	70	8000	None

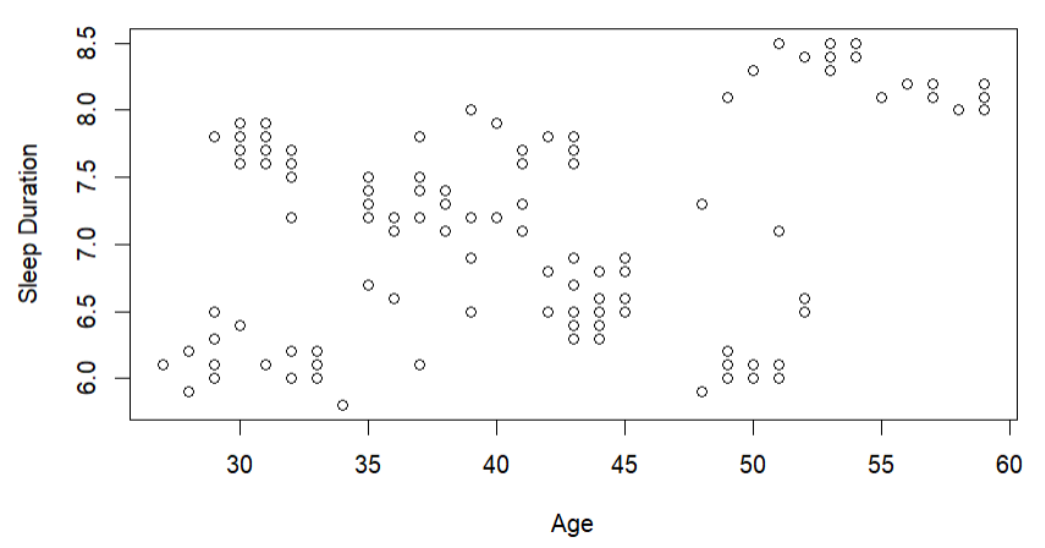
SUMMARY STATISTICS

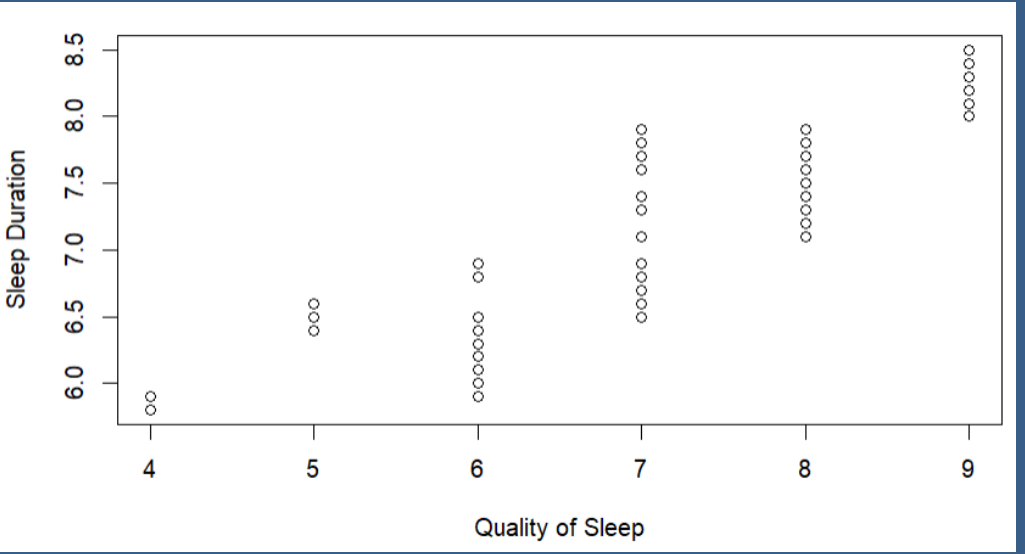
```
      Age        Sleep.Duration  Quality.of.Sleep Physical.Activity.Level  Stress.Level     Heart.Rate
 Min.   :27.00   Min.   :5.800   Min.   :4.000    Min.   :30.00           Min.   :3.000   Min.   :65.00
 1st Qu.:35.25   1st Qu.:6.400   1st Qu.:6.000    1st Qu.:45.00           1st Qu.:4.000   1st Qu.:68.00
 Median :43.00   Median :7.200   Median :7.000    Median :60.00           Median :5.000   Median :70.00
 Mean   :42.18   Mean   :7.132   Mean   :7.313    Mean   :59.17           Mean   :5.385   Mean   :70.17
 3rd Qu.:50.00   3rd Qu.:7.800   3rd Qu.:8.000    3rd Qu.:75.00           3rd Qu.:7.000   3rd Qu.:72.00
 Max.   :59.00   Max.   :8.500   Max.   :9.000    Max.   :90.00           Max.   :8.000   Max.   :86.00

  Daily.Steps       Occupation    Gender       Sleep.Disorder        BMI.Category Blood.Pressure
 Min.   : 3000   Nurse      :73   Female:185   Insomnia   : 77   Normal       :195   Min.   :0.0000
 1st Qu.: 5600   Doctor     :71   Male  :189   None       :219   Normal Weight: 21   1st Qu.:0.0000
 Median : 7000   Engineer   :63                Sleep Apnea: 78   Obese        : 10   Median :1.0000
 Mean   : 6817   Lawyer     :47                                  Overweight   :148   Mean   :0.5561
 3rd Qu.: 8000   Teacher    :40                                                      3rd Qu.:1.0000
 Max.   :10000   Accountant:37                                                      Max.   :1.0000
                 (Other)    :43
```

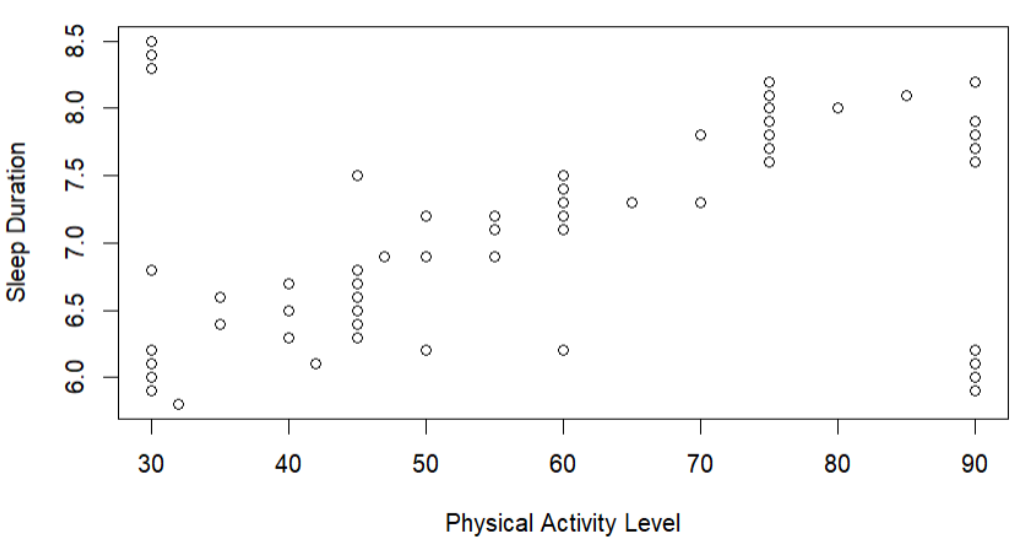
The response variable for this dataset, Duration of Sleep, has a range of values is between 5.800 and 8.500, with a mean of 7.130.

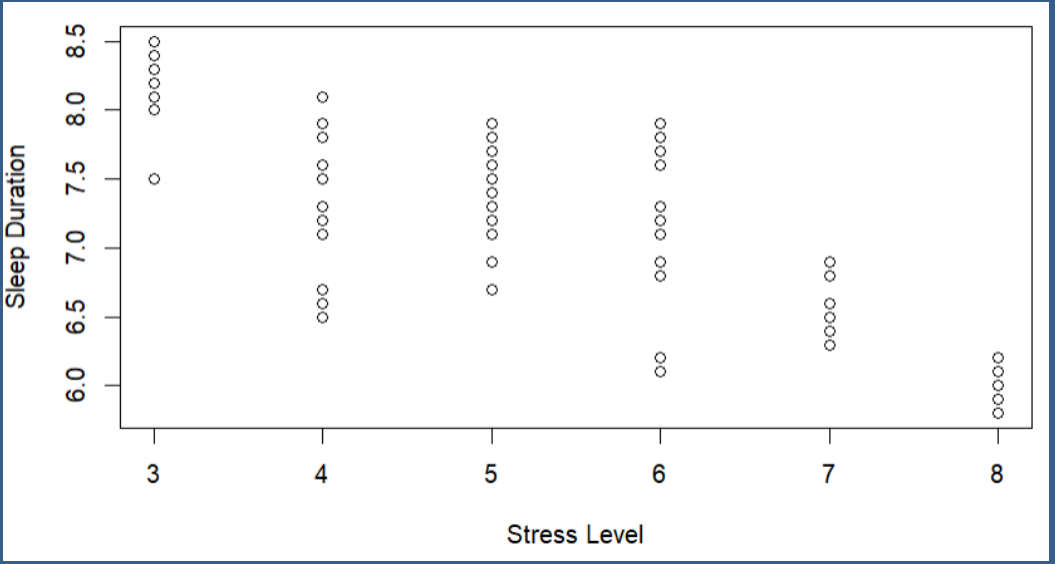


UNIVARIATE PLOTS

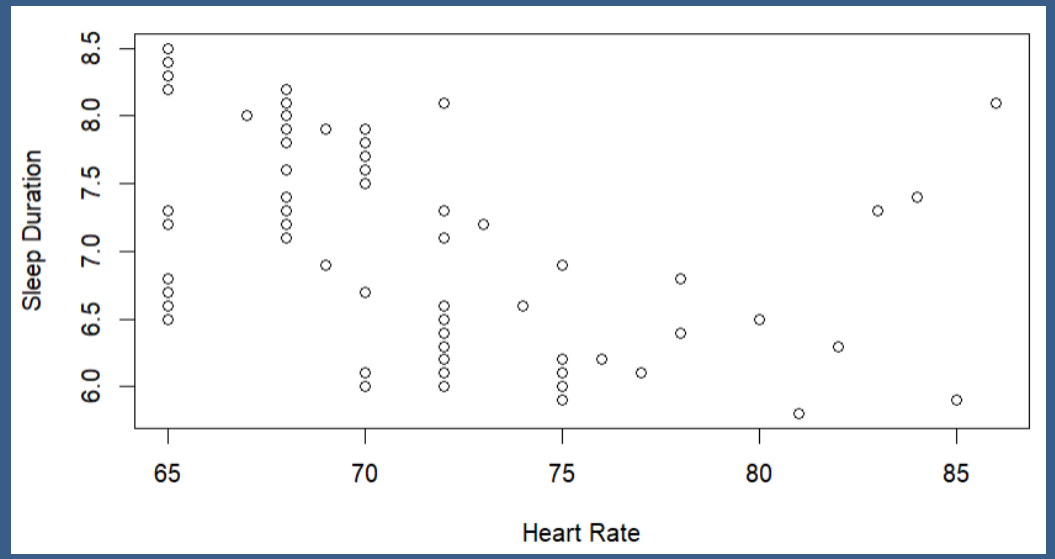


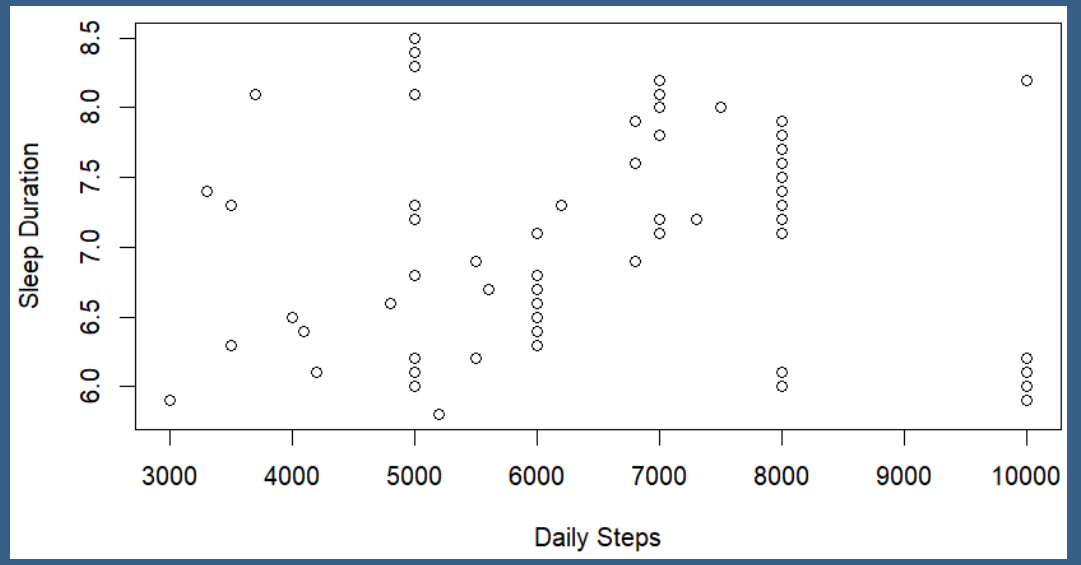


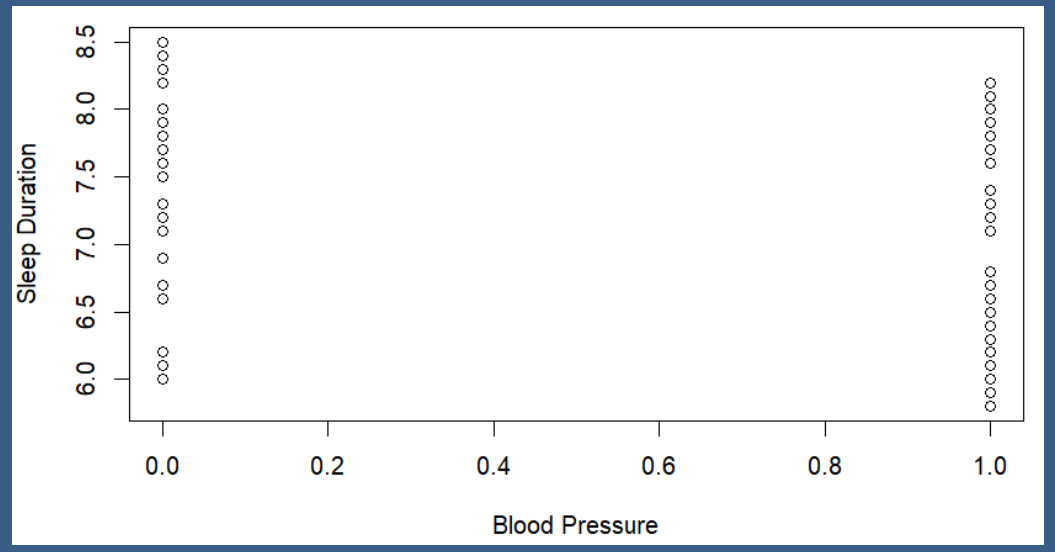


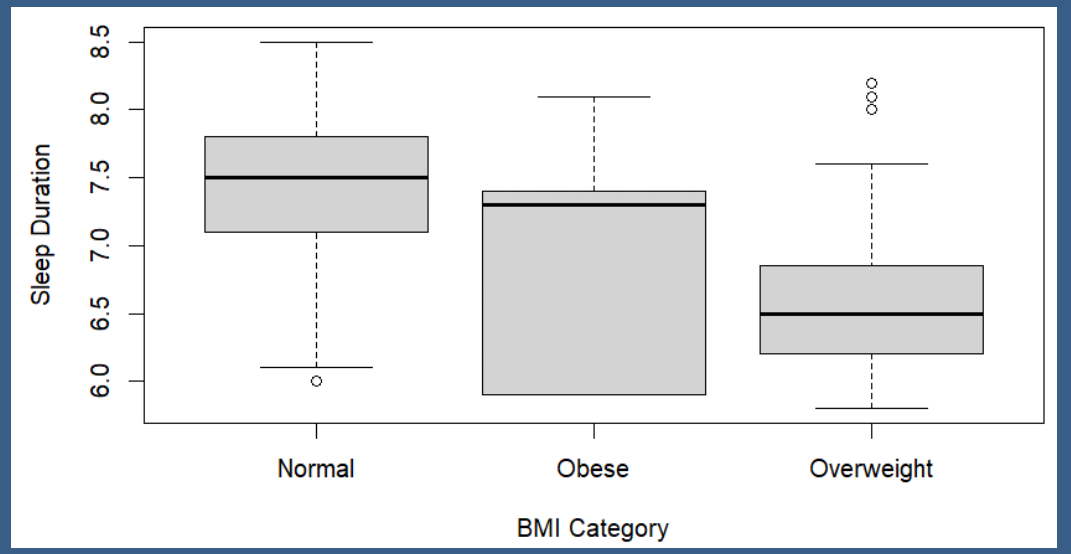


UNIVARIATE PLOTS

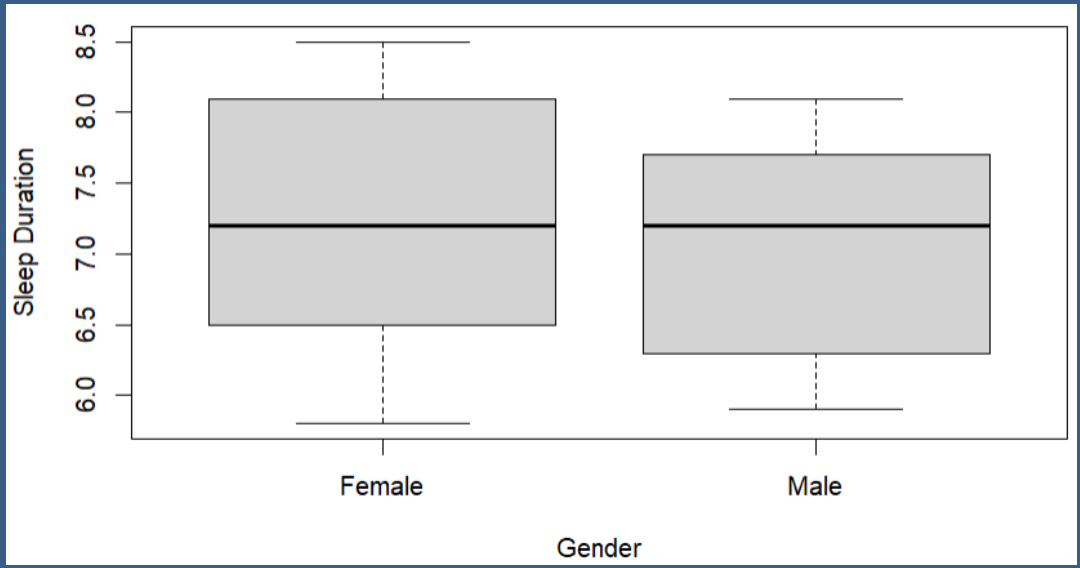


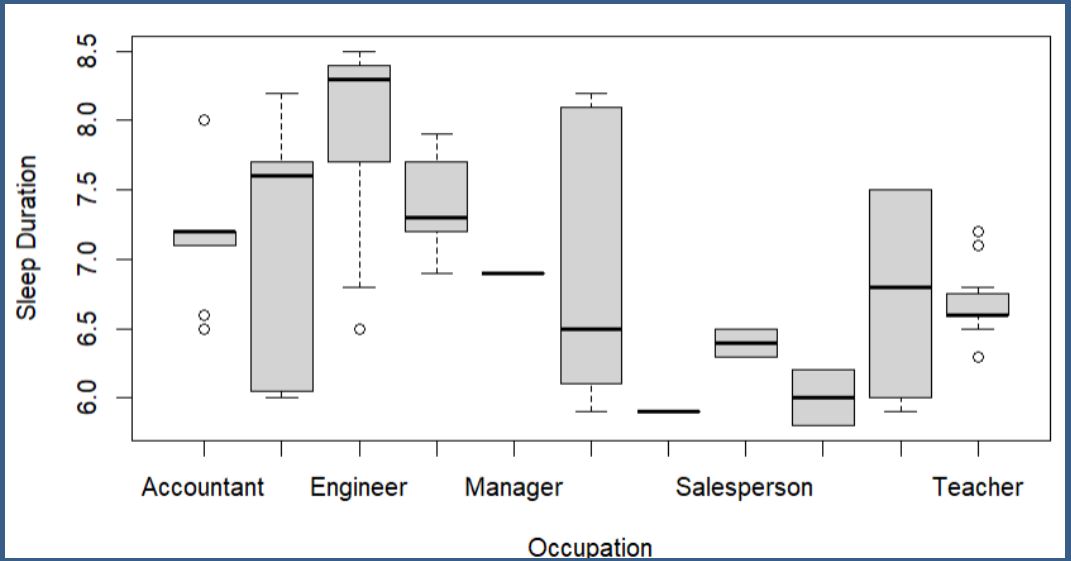


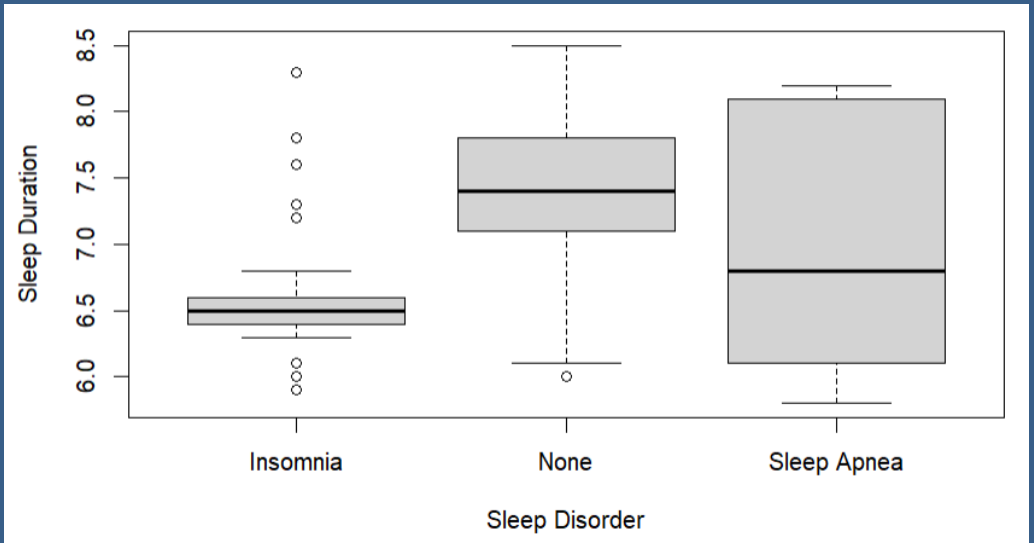




UNIVARIATE PLOTS







CLEANING DATA

BMI

- Some of the observations in the BMI Category column had a listing of “Normal”, while others had “Normal Weight” instead.
- These seemed to be same category and therefore all “Normal Weight” observations were changed to “Normal” for consistency.

MISSING VALUES

- No missing values were in the data to fix.

FULL MODEL

$$Y = \beta_1(\text{GenderMale}) + \beta_2(\text{Age}) + \beta_3(\text{OccupationDoctor}) + \beta_4(\text{OccupationEngineer}) + \beta_5(\text{OccupationLawyer}) + \beta_6(\text{OccupationManager}) + \beta_7(\text{OccupationNurse}) + \beta_8(\text{OccupationSalesRepresentative}) + \beta_9(\text{OccupationSalesperson}) + \beta_{10}(\text{OccupationScientist}) + \beta_{11}(\text{OccupationSoftwareEngineer}) + \beta_{12}(\text{OccupationTeacher}) + \beta_{13}(\text{Quality.of.Sleep}) + \beta_{14}(\text{Physical.Activity.Level}) + \beta_{15}(\text{Stress.Level}) + \beta_{16}(\text{BMI.CategoryObese}) + \beta_{17}(\text{BMI.CategoryOverweight}) + \beta_{18}(\text{Heart.Rate}) + \beta_{19}(\text{Daily.Steps}) + \beta_{20}(\text{Sleep.DisorderNone}) + \beta_{21}(\text{Sleep.DisorderApnea}) + \beta_{22}(\text{Blood.Pressure}) + E$$

- 11 Independent Variables (k = 11)
- Gender has 1 dummy variable
- Occupation has 10 dummy variables
- BMI Category has 2 dummy variables
- Sleep Disorder has 2 dummy variables

LINEAR REGRESSION ASSUMPTIONS

REGRESSION ASSUMPTIONS

EXISTENCE

Existence is valid as each x value for all the variables has a possible random Y variable. All regression models meet this assumption.

INDEPENDENCE

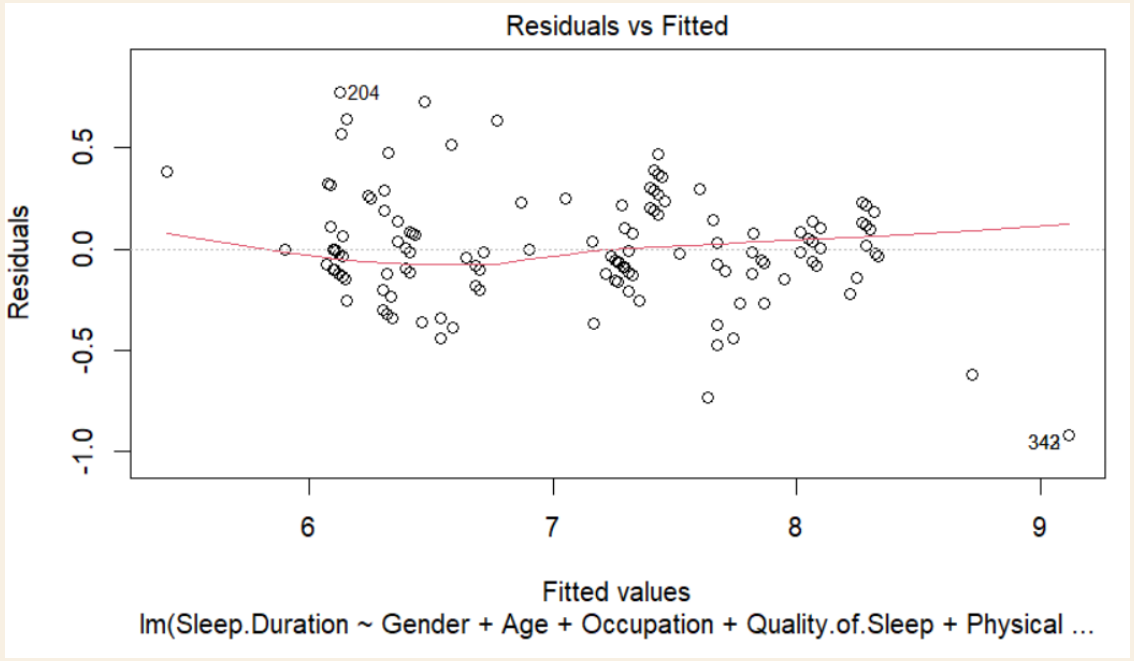
Independence is validated as each observation was made for a unique individual, as noted by their unique ids in the first column of the data set.

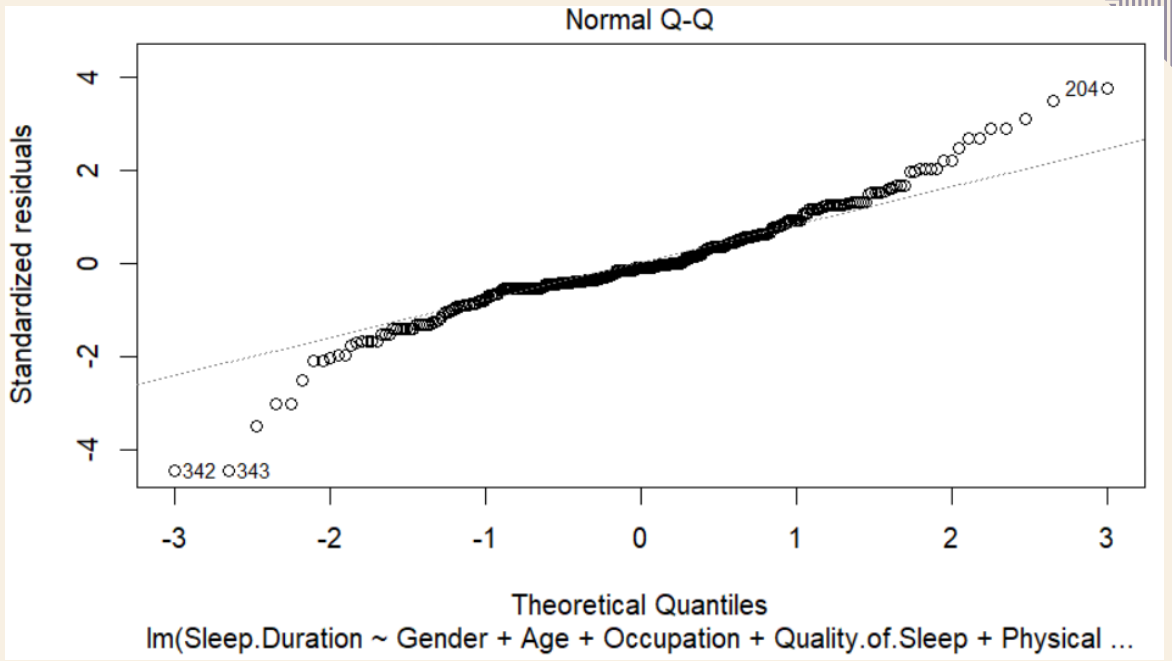
LINEARITY

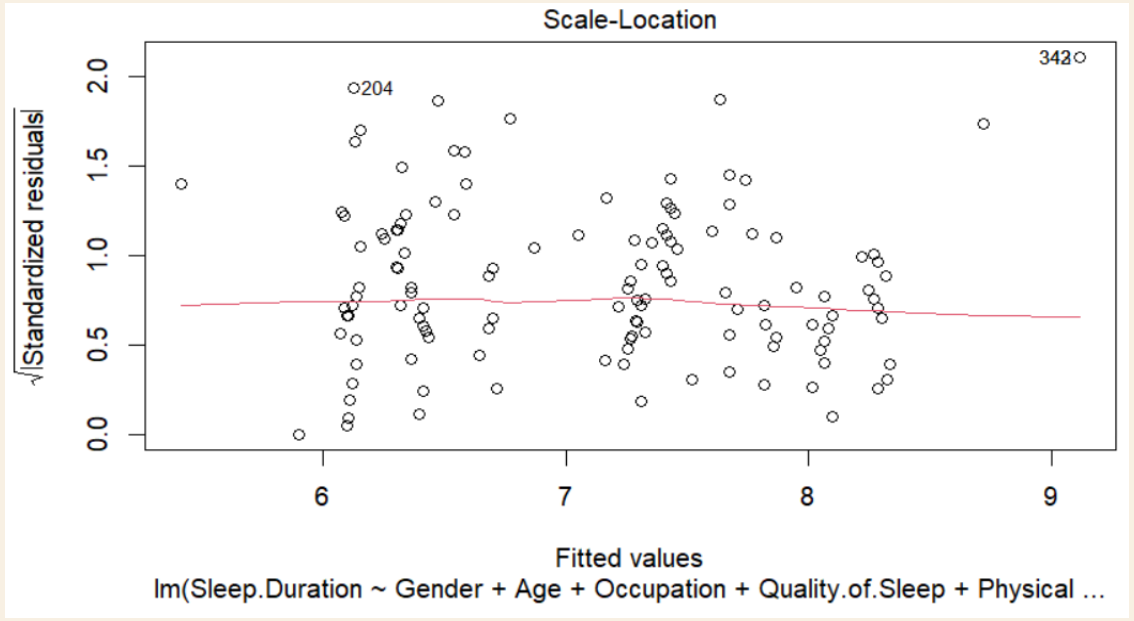
The assumption most questionable for our dataset.

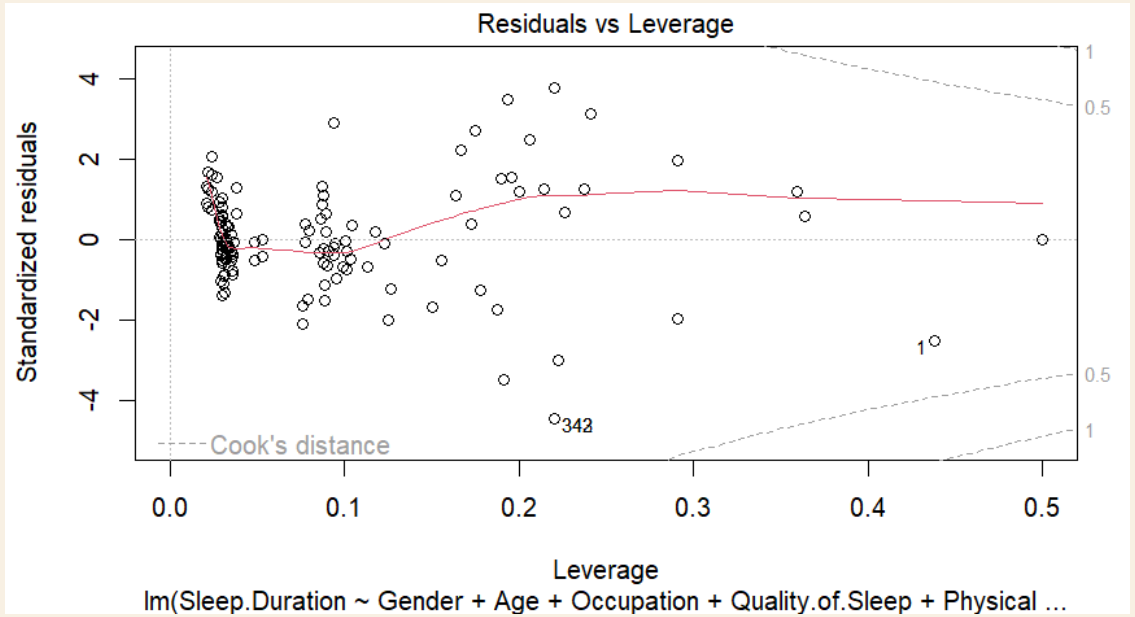
Most of the independent variables show a linear relationship through the individual scatterplots/box plots. However, a few of the variables did seem to have a non or curlinear relationship that could violate this assumption.

REGRESSION ASSUMPTIONS - GRAPHS





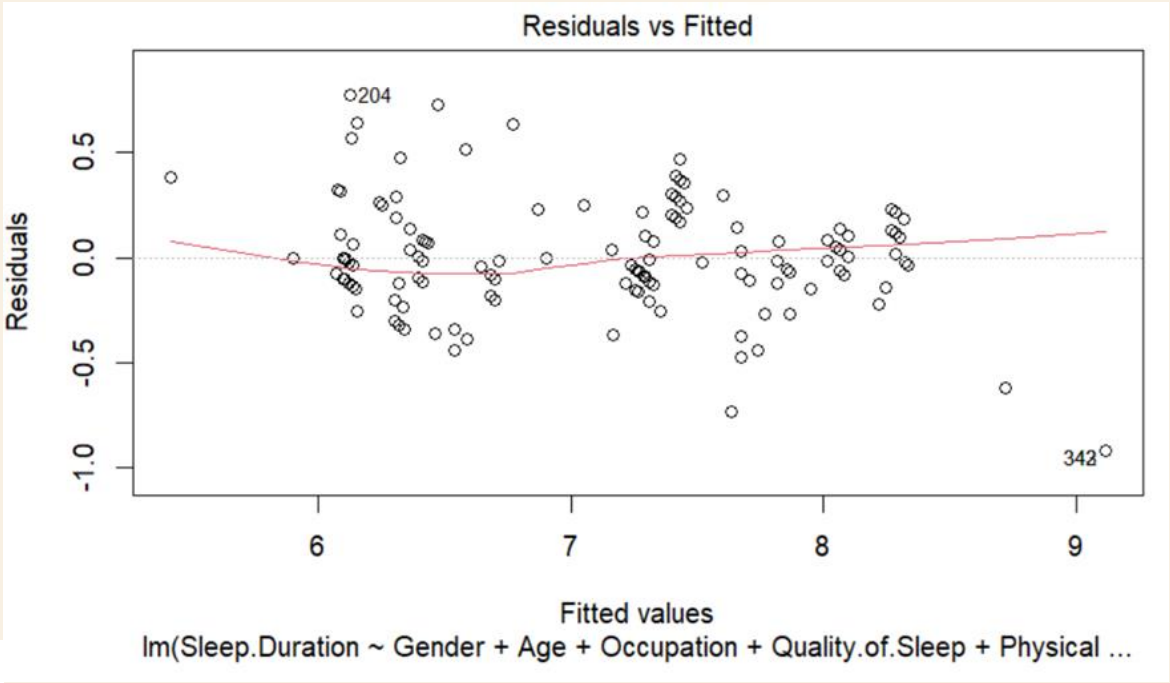


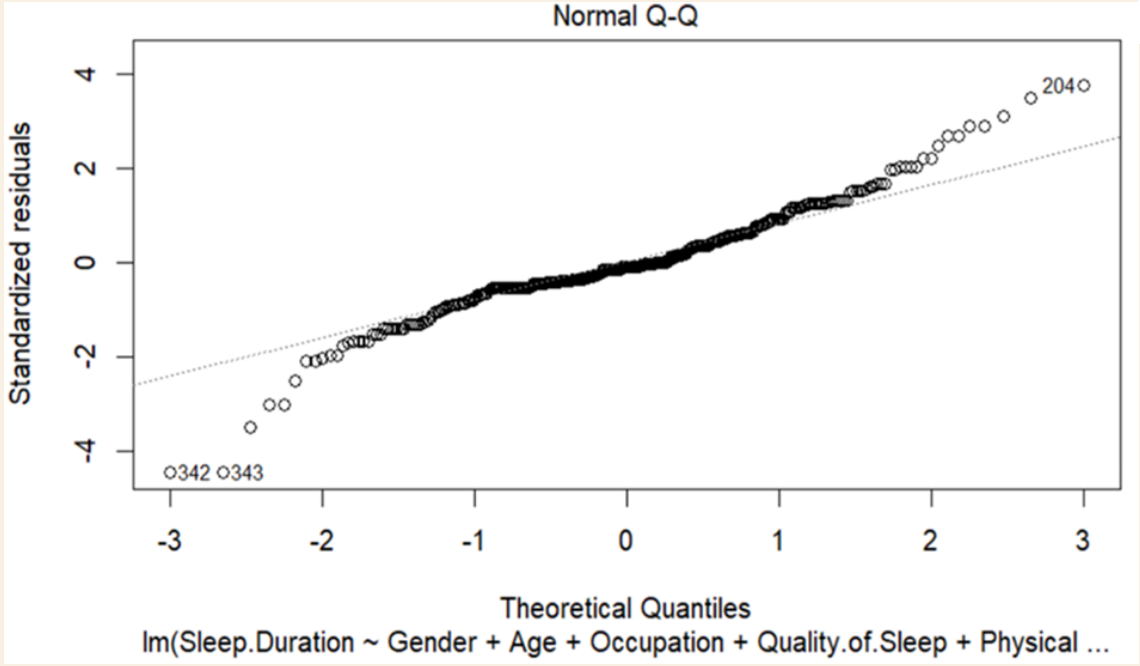


REGRESSION ASSUMPTIONS - GRAPHS

RESIDUAL VS FITTED GRAPH

- The even spread of the residual points validates Homoscedasticity, implying a constant variance
- The relatively linear fitted line of the graph also adds credit to linearity of the entire model.



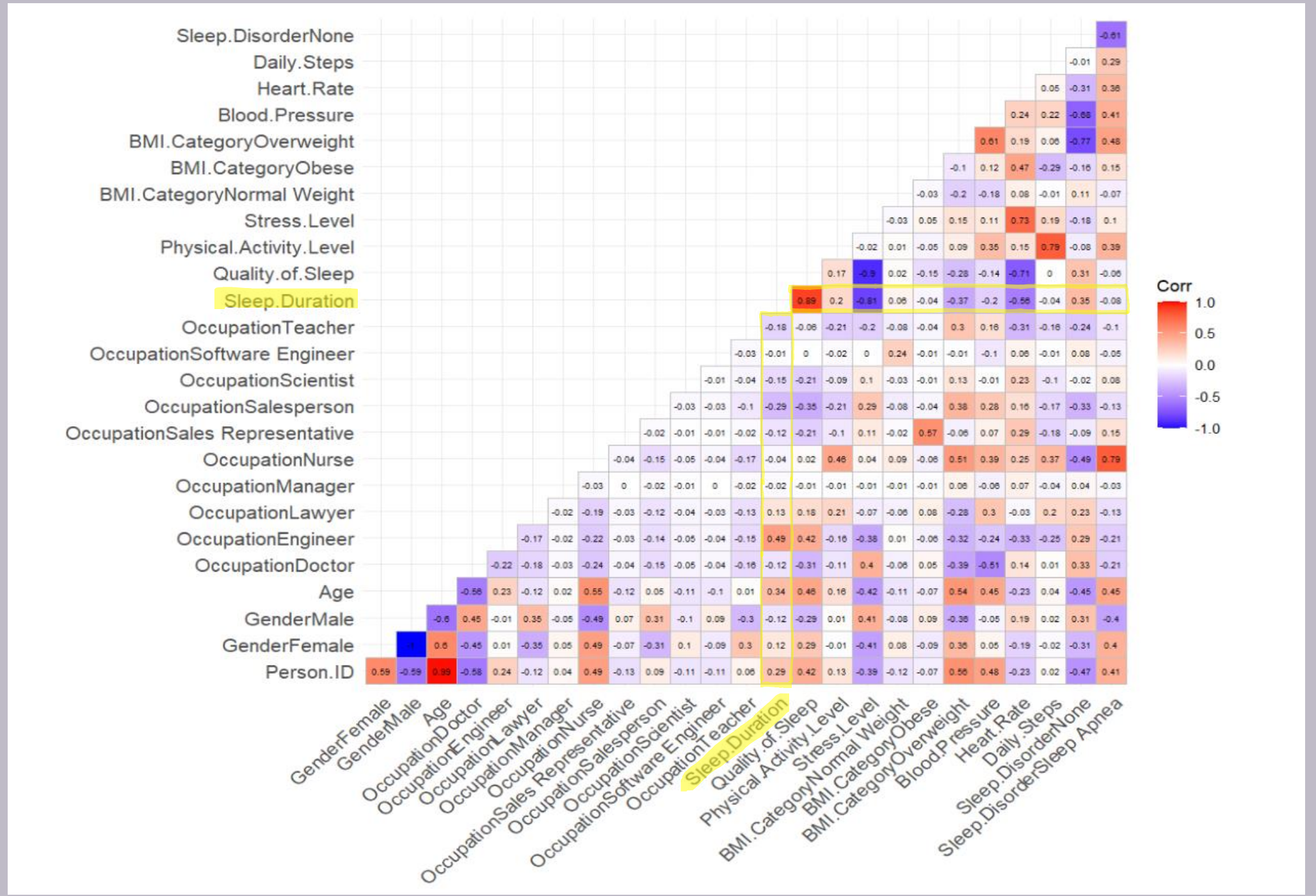


Q-Q PLOT

- Despite some trailing at the tails, the dots follow a straight line well, meaning the model follows a normal distribution, validating the normality assumption.

COLLINEARITY

- In our exploration of the data, collinearity was noticed between multiple variables.
- We looked for instances of collinearity through two methods:
  - Correlation Matrix
  - Variance Inflation Factor (VIF) tests for each predictor
- Through these methods, especially the VIF tests, we will remove three independent variables from our model: Quality of Sleep, Heart Rate, and Gender.


Sleep.DisorderNone
Daily.Steps
Heart.Rate
Blood.Pressure
BMI.CategoryOverweight
BMI.CategoryObese
BMI.CategoryNormal Weight
Stress.Level
Physical.Activity.Level
Quality.of.Sleep
Sleep.Duration
OccupationTeacher
OccupationSoftware Engineer
OccupationScientist
OccupationSalesperson
OccupationSales Representative
OccupationNurse
OccupationManager
OccupationLawyer
OccupationEngineer
OccupationDoctor
Age
GenderMale
GenderFemale
Person.ID
GenderFemale
GenderMale
Age
OccupationDoctor
OccupationEngineer
OccupationLawyer
OccupationManager
OccupationNurse
OccupationSales Representative
OccupationSalesperson
OccupationScientist
OccupationSoftware Engineer
OccupationTeacher
Sleep.Duration
Quality.of.Sleep
Physical.Activity.Level
Stress.Level
BMI.CategoryNormal Weight
BMI.CategoryObese
BMI.CategoryOverweight
Blood.Pressure
Heart.Rate
Daily.Steps
Sleep.DisorderNone
Sleep.DisorderSleep Apnea
Corr
1.0
0.5
0.0
-0.5
-1.0

COLLINEARITY – VIF TESTS

FIRST TEST

- There was collinearity shown in Gender, Age, Stress Level, BMI, Heart Rate, and Quality of Sleep.
- Quality of Sleep was removed for having the largest VIF.

Variables <chr>	Tolerance <dbl>	VIF <dbl>
GenderMale	0.06857103	14.583419
Age	0.05390193	18.552211
OccupationDoctor	0.11220746	8.912064
OccupationEngineer	0.13657790	7.321829
OccupationLawyer	0.16594509	6.026090
OccupationManager	0.86975885	1.149744
OccupationNurse	0.14413883	6.937756
OccupationSales Representative	0.47843194	2.090161
OccupationSalesperson	0.15518398	6.443964
OccupationScientist	0.55520873	1.801124
OccupationSoftware Engineer	0.72845283	1.372772
OccupationTeacher	0.24840135	4.025743
Physical.Activity.Level	0.15129285	6.609697
Stress.Level	0.04006199	24.961315
BMI.CategoryObese	0.18627748	5.368335
BMI.CategoryOverweight	0.09453864	10.577685
Heart.Rate	0.09428547	10.606088
Quality.of.Sleep	0.03055791	32.724754
Daily.Steps	0.16701112	5.987625
Sleep.DisorderNone	0.16692186	5.990827

SECOND TEST

- After removing Quality of Sleep, most of the VIFs went significantly down.
- Heart Rate and Gender still colinear, removed Heart Rate next.

Variables <chr>	Tolerance <dbl>	VIF <dbl>
Age	0.1959201	5.104121
OccupationDoctor	0.3266605	3.061282
OccupationEngineer	0.3408506	2.933837
OccupationLawyer	0.2846272	3.513367
OccupationManager	0.9209907	1.085787
OccupationNurse	0.1661700	6.017934
OccupationSales Representative	0.6066004	1.648532
OccupationSalesperson	0.3214120	3.111271
OccupationScientist	0.6936811	1.441585
OccupationSoftware Engineer	0.8923413	1.120647
OccupationTeacher	0.3486152	2.868492
Physical.Activity.Level	0.2356914	4.242836
Stress.Level	0.3321933	3.010296
BMI.CategoryNormal Weight	0.7385652	1.353977
BMI.CategoryObese	0.4608972	2.169681
BMI.CategoryOverweight	0.1337794	7.474995
Daily.Steps	0.2138371	4.676458

THIRD TEST

- All variables except Gender are now under 10 VIF.
- We remove Gender.
- One final test after shows all the VIFs are now under 10.

INTERACTION TERMS

- Looking at the data there were variables we suspected of interaction.
- To test for interactions, we did a regression of each suspected interactions with sleep duration to see if the interaction term was statistically significant. If it was, we added it to the model.
- We found two significant interactions: Physical Activity and Daily Steps as well as Physical Activity and Age. Therefore, both were added to the model.

Example of Testing Interaction for Physical Activity and Age:

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.7466437  0.4729110  -1.579    0.115
Physical.Activity.Level    0.1141811  0.0077171  14.796   <2e-16 ***
Age                        0.1748728  0.0108388  16.134   <2e-16 ***
Physical.Activity.Level:Age -0.0024748  0.0001727 -14.327   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

MODEL SELECTION

MAXIMUM MODEL

- $Y = \beta_1(Age) + \beta_2(OccupationDoctor) + \beta_3(OccupationEngineer) + \beta_4(OccupationLawyer) + \beta_5(OccupationManager) + \beta_6(OccupationNurse) + \beta_7(OccupationSalesRepresentative) + \beta_8(OccupationSalesperson) + \beta_9(OccupationScientist) + \beta_{10}(OccupationSoftwareEngineer) + \beta_{11}(OccupationTeacher) + \beta_{12}(Physical.Activity.Level) + \beta_{13}(Stress.Level) + \beta_{14}(BMI.CategoryObese) + \beta_{15}(BMI.CategoryOverweight) + \beta_{16}(Daily.Steps) + \beta_{17}(Sleep.DisorderNone) + \beta_{18}(Sleep.DisorderApnea) + \beta_{19}(Blood.Pressure) + \beta_{20}(Physical.Activity.Level*Daily.Steps) + \beta_{21}(Physical.Activity.Level*Age) + E$

Three Selection Methods were used:

- Backwards Elimination
- Forwards Selection
- Stepwise Variable Selection

BACKWARDS ELIMINATION

- Backwards Elimination using p-value eliminated one variable: Blood Pressure.

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Blood.Pressure	0.9446	0.9415	0.2181	-147.3045	0.1924

FORWARD-SELECTION

- Forward-Selection eliminated none of the variables.

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Stress.Level	0.6543	0.6533	1821.0791	492.4490	0.4682
2	Occupation	0.8518	0.8472	573.6550	199.0811	0.3108
3	Physical.Activity.Level	0.8754	0.8712	426.4007	136.9321	0.2854
4	Age	0.8792	0.8748	404.3879	127.4876	0.2814
5	Age:Physical.Activity.Level	0.9221	0.9190	134.7226	-32.9901	0.2263
6	Sleep.Disorder	0.9259	0.9225	112.8710	-47.3490	0.2213
7	Daily.Steps	0.9280	0.9245	101.4273	-56.1129	0.2185
8	BMI.Category	0.9393	0.9360	31.7827	-115.4419	0.2011
9	Blood.Pressure	0.9404	0.9370	26.9944	-120.0450	0.1996
10	Physical.Activity.Level:Daily.Steps	0.9450	0.9417	0.0000	-147.6553	0.1921

STEPWISE VARIABLE SELECTION

- The Stepwise Selection Method eliminated only Blood Pressure. Same model as Backward Elimination's model.

Backward Elimination Summary

Variable	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
Full Model	-209.458	10.803	222.529	0.95370	0.95077
Blood.Pressure	-211.056	10.814	222.517	0.95365	0.95086

COMPARING METHOD RESULTS

Backwards & Stepwise

Parameters = 9

R^2 = 0.9446

MSE = 0.1924

Frontward Selection

Parameters = 10

R^2 = 0.9450

MSE = 0.1921

Verdict

The Model from the Backwards & Stepwise Methods is chosen. While the R^2 and MSE of both models are nearly identical, the Backwards and Stepwise model has one less parameter.

FINAL MODEL

- $Y = \beta_1(\text{Age}) + \beta_2(\text{OccupationDoctor}) + \beta_3(\text{OccupationEngineer}) + \beta_4(\text{OccupationLawyer}) + \beta_5(\text{OccupationManager}) + \beta_6(\text{OccupationNurse}) + \beta_7(\text{OccupationSalesRepresentative}) + \beta_8(\text{OccupationSalesperson}) + \beta_9(\text{OccupationScientist}) + \beta_{10}(\text{OccupationSoftwareEngineer}) + \beta_{11}(\text{OccupationTeacher}) + \beta_{12}(\text{Physical.Activity.Level}) + \beta_{13}(\text{Stress.Level}) + \beta_{14}(\text{BMI.CategoryObese}) + \beta_{15}(\text{BMI.CategoryOverweight}) + \beta_{16}(\text{Daily.Steps}) + \beta_{17}(\text{Sleep.DisorderNone}) + \beta_{18}(\text{Sleep.DisorderApnea}) + \beta_{19}(\text{Physical.Activity.Level*Daily.Steps}) + \beta_{20}(\text{Physical.Activity.Level*Age}) + E$

FINAL THOUGHTS

OVERFITTING?

- The very high R^2 of 0.95 and very low MSE of 0.19 for the final model may suggest overfitting of the model.

RELIABILITY

- Due to earlier linearity assumptions being questionable for some of the predictor variables, the model may not be very reliable.
- Furthermore, if overfitting is true, that would further cast doubt on the model's reliability on other data.

NEXT STEPS

- Next could be to run tests to check for overfitting, likely through rebuilding the model with cross validation / split sampling.

THANK YOU

Ava Nafisi
904-514-1798
av549843@ucf.edu