Project Big Data

Assignment 3 Report

Yasin Aydin yasinaydin@sabanciuniv.edu
Faruk Simsekli faruksimsekli.7@gmail.com

Note: When we finished the assignment the new version of assignment 3 was not released so that the following quote is not follow in the same way.“Remove the rows that have NaN values in this merged DataFrame.”. Still we managed to not use those rows with NaN values. So, the result will not change at the end. Also, document is a bit longer than maximum size since we already completed before the limit announcement.

Correlation results:

Some correlation results between values from the datasets.

- The correlation between **bedtime procrastination scale** (bp_scale, a personality trait) and **mean** time spent **delaying bedtime** (using Pearson correlation test) is **-0.375**, which shows an inverse correlation. (p-value = 0.020)

- The correlation between **age** and **mean** time spent **delaying bedtime** (using Kendall rank correlation) is **-0.136**, which shows the small inverse correlation. (p-value = 0.243)

- The correlation between **mean** time spent **delaying bedtime** and **daytime sleepiness** (using Pearson correlation test) is **0.020**, which shows that there is almost no correlation. (p-value = 0.906)

Experimental group vs. Control group:

To determine significant differences between the experimental group and control group. We have checked some of the values from the data set to detect those differences.

When comparing the first values we used **wilcoxon rank-sum test** because the value are not continues which implies that we can not use **t-test**. When comparing the second and third values we used **ttest_ind**, which calculates the means of two independent samples of scores. The reasons are the following. Firstly, we assumed that people are chosen independently from the **same distribution**, meaning that these 2 samples have identical expected and variance values.

Results:

1. The difference of the number of nights participants delayed their bedtime between experimental and control group is **0.329**. (p-value = 0.743) - (**wilcoxon**)

2. The difference of the time participants spent in bed each night between experimental and control group is **0.372**. (p-value = 0.712) - (**t-test**)

3. The difference of the mean time participants spent delaying their bedtime between experimental and control group is **-1.790**. (p-value = 0.082) - (**t-test**)

The **probability value** tells us what the odds are that our results could have happened by chance. Most of the **probability values** that are calculated, were higher than 5% (except the first one), which supports the null hypothesis. The **null hypothesis** is a general statement that there is no relationship between two measured phenomena, or no association among groups. So, with the p-values we can conclude that there is no relationship between values from the experimental group and control group.
Also, the **t-statistic** gives us the evidence of how much experimental group and control group differs from each other. The results are **0.364, 0.372, -1.79**. These t-statistics can not help us deduct strong arguments since p-values are bigger than **5% (0.05)**.

Hypothesis:

The age and motivation is inversely but the daytime sleepiness is directly proportional to people's bedtime delay time.

We think that younger people tend to delay their bedtimes with the 21st century technology. Also, older people tend to have sleep patterns which makes it harder to delay bedtime in a daily basis. We think that motivation gives people to start next day early so that they tend to plan and not delay bedtimes. We also think that people with high bedtime delay times tend to sleep less, assuming that they have to get up at the same hours everyday. This makes their daytime sleepiness higher which makes us to add this as a factor.

Results of regression model:

1 - Coefficients:

We have found the coefficients of the factors age, motivation, daytime_sleepiness respectively as follows: **-27.31**, **446.58**, and **104.53**. This clearly reveals that even though the age is inversely proportional to bedtime delay. On the other hand, motivation and daytime_sleepiness are directly proportional to bedtime delay. The results show that in our hypothesis we were wrong about the relationship between motivation and bedtime delay because we predicted that motivation is also inversely proportional to bedtime delay. However, according to our analysis, this is not the case. Moreover, we have seen that absolute value of the factor age's coefficient is the smallest. We think that this is related to the scale of the factor. Whereas the scale of the factor age is 18 to 60, the scale of the other factor is less.

2 - Standard Error:

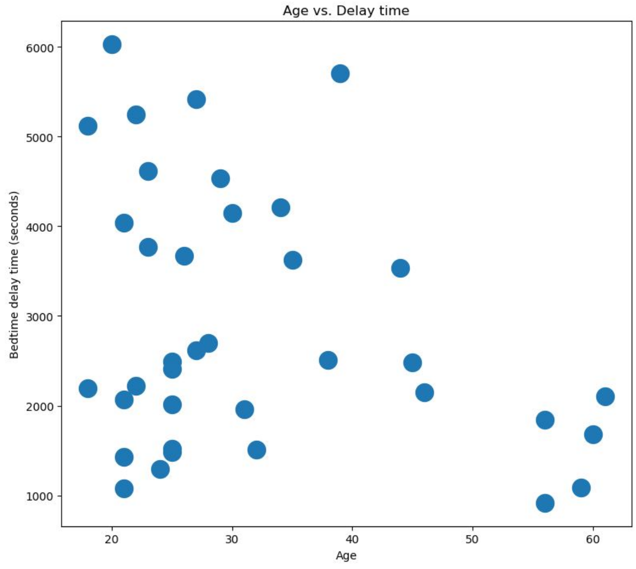
The standard error of age factor is **19.4**, which is a small number when compared to big delay times of the dataset. The standard error of motivation factor is **192.5**, which is 10 times bigger than age factor. This value shows that the motivation factor is not that accurate to use it as a determinant factor. The standard error of daytime sleepiness is **51.1**. This value is also small when compared to the scale of delay time.

To conclude age and sleepiness factors can be used to predict more accurate results but the standard error is not good enough by itself. We need to consider p-value and other values as well.

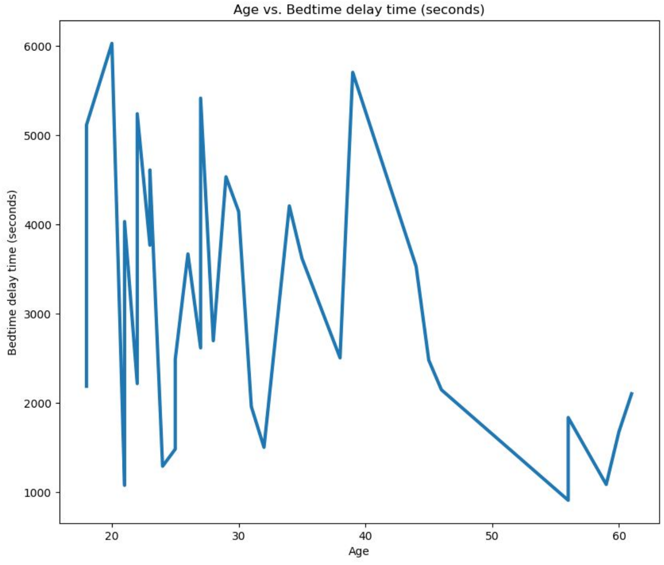
3 - Probability value:

The p-value of age factor is **0.169** and this value indicates weak evidence against the null hypothesis, so it fails to reject the null hypothesis. The p-value of motivation factor is **0.026**. The p-value of daytime sleepiness factor is **0.049**. So, motivation and daytime sleepiness factors are below 5% which means that they reject the null hypothesis. This also means that they affect the bedtime delay time.

Visualization:



Graph 1.



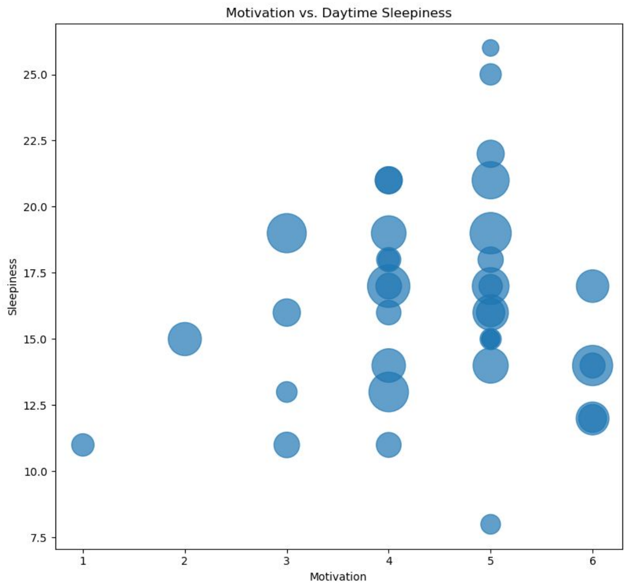
Graph 2.

(Graph 2 was created to make visualization of the slope easier)

Graph 1 above shows the Age and Delay time values. With this data we can say that old people (50+) in the population tend to not delay their bedtimes as much as young people(50>).

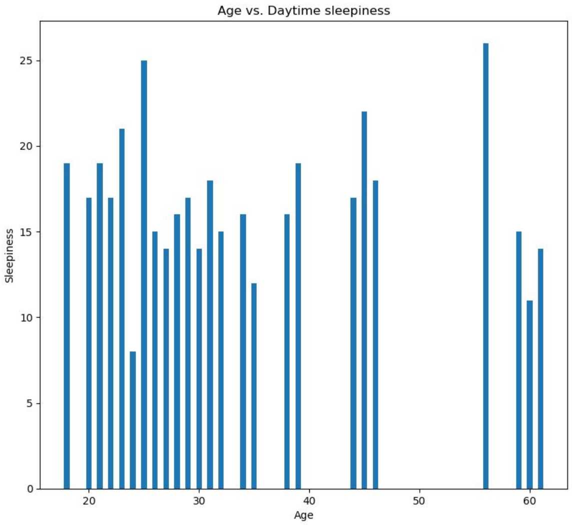
Also, the highest delay times are generally between 18 to 30 age group. These graphs show that the age part of our hypothesis and the reasons we came up with for it were true. Yet, above we've seen that p-value of age is too high to affect the delay time. We think that this contradiction is because we do not have enough sample. Graphs would have more uniform distribution if the top right part of it was filled with some examples but since there is no examples data that has delay time more than 3000 and age more than 45 we could come up with this deductions.

The cluster that is located at the bottom left side in graph 1 affects the statistics so that there are no correlation between age and bedtime delay time but as seen in the graph we can correlate age factor with the bedtime delay time. This might be again related to the number of samples we have. Also, the relation that age affects the delay time inversely, in graph 2 *is seen more clear than graph 1.*

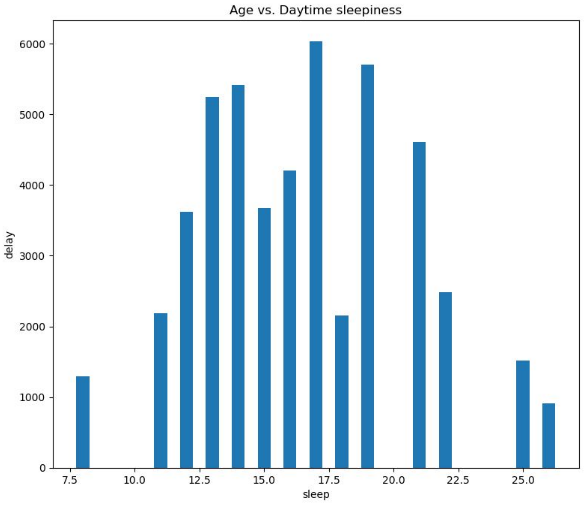


Graph 3.

The graph above shows the relation between motivation, daytime sleepiness and the delay times with the bubble size. This graph disproves second part of our hypothesis, which was that the motivation values are inversely proportional to the delay times of people. In the graph, it can be clearly seen that people who had ‘6’ motivation points have high delay times. We can say that people with ‘6’ motivation point have less sleepiness but the relation between factor does not affect the delay time results since there is no direct or inverse proportion between sleepiness and delay time. The bubble sizes and the values of sleepiness is not giving us any pattern that we can use. Furthermore, we realized that people who has the more delay time are the ones who are in the range (12-21) of daytime sleepiness.



Graph 4



Graph 5

In *graph 4*, we can see that most of the people from any age group have sleepiness between 10 and 20 which means that we can not correlate this values and deduce something from them. However, in the *graph 5*, we can see that people with very low and very high sleepiness has low bedtime delay times. In the *graph 5*, there is a kind of triangle shape that can be seen from the corners to the middle top side but this does not show that our hypothesis is correct since there isn’t any clear structure. Also, we thought that there is a direct proportion between daytime sleepiness and bedtime delay time but it seems like *graph 5* disproves that.

Conclusion:

We thought that age and motivation is inversely proportional to the bedtime delay time. The regression model showed that the age can not be correlated with the delay time (since the p-value is **0.169**) but with the graphs we acquired showed that there was a clear inverse proportion. These results probably conflict because of the lack of samples. On the other hand, the motivation factor has **0.026** p-value which shows that it is affecting the bedtime delay time but the coefficient is not negative so the proportion is not inverse as we thought. We also thought that daytime sleepiness is directly proportional to the bedtime delay time and the **0.049** p-value with the positive coefficient showed that our assumption is correct.

References

https://docs.scipy.org/doc/scipy/reference/stats.html
https://www.statisticshowto.datasciencecentral.com/t-statistic/
https://www.statisticshowto.datasciencecentral.com/p-value/
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/null-hypothesis/
https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687
https://datatofish.com/statsmodels-linear-regression/