**Data Preparation:**

After the basic data exploration, we realized that we need to change our data to be numeric, since our model doesn’t know how to deal with English letters, only numbers.

* First, we changed the ‘gender’ columns from male or female to 0 or 1 accordingly.
* Second, we changed the ‘smoking’, ‘heart\_disease\_hist’, ‘heart\_disease\_family\_hist’, ‘bp\_medication’ and ‘diabetes’ columns from Yes or No to 0 or 1 accordingly.
* Third, we changed ‘work\_stress\_level’ and ‘exercise\_level’ columns from X/5 to X\*0.20.

After doing so, since each measure have different scale, we needed to normalize the results.

For example: ‘heart\_disease\_hist’ column was at that point 0 or 1 for each person. This is obviously a very important indicator for the chance of hurt flier1, but the deference between having, it or not is only 1 unit. However, the ‘height’ column had values between 150 – 195 so the difference can be up to 45 units, but it doesn’t tell us much about the health of the person.

So, at first, we tried to create formulas that will create a binary (0/1) form to every relevant column:

* ‘bmi\_bit’ = if normal BMI (18.5-25) add 1, else, 0. Excel formula: (IF(AND(bmi>18.5,bmi<25),1,0))
* ‘Ecg\_score\_bit’ = Since more than 75% of people got score 100/100 appendix 1 and there is no scale like that online (probably its PreSee scale) than we dicided that if its 100 than add 1, else add 0. Excel formula: IF(acg=100,1,0)
* ‘us\_test\_score\_bit’ = Same as the Ecg\_score\_bit appendix 2. Excel formula : IF(us\_test\_score=100, 1, 0)
* ‘blood\_test\_score\_bit’ = If the score was higher (better) then median score (=87) appendix 3
* ‘bp\_systolic\_bit’ = If the systolic bp was too high (> 125)2  add 0, else 1.  
  Excel formula : IF(bp\_systolic > 125, 0, 1)
* ‘bp\_diastolic\_bit’ = If the diastolic bp was too low (< 80) 2 add 0, else 1.  
   Excel formula : IF(AE2<80, 1, 0)

Then we ran our model (that we will talk more about in the next section) with the binary data.   
We noticed a problem. The Kmean algorithm we use didn’t gave us useful results. The algorithm kept splitting the data to similar sizes groups in a linear way as long as we kept increasing the K. In addition there wasn’t a meaningful difference between the groups (mostly in the first 2 columns - ‘gender’ and ‘age\_Bit’).

Here are the results from this try:

1. [Family History of Coronary Heart Disease, a Strong Risk Factor for Myocardial Infarction Interacting with Other Cardiovascular Risk Factors: Results from the Stockholm Heart Epidemiology Program (SHEEP)](https://www.jstor.org/stable/3703625), *by Karin L., Johan H., Christina R. and Anders A,* **Epidemiology** Vol. 12, No. 2 (Mar., 2001), pp. 215-221 (7 pages)
2. <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>

Chart, line chart

Description automatically generatedTable

Description automatically generatedChart, line chart

Description automatically generated

We understood that too much information was lost while taking scaled information and making it binary. Since the scale was lost, the model didn’t have enough information to work properly.

So, instead of using the binary formed information we used normalization formula:  
**(cell\_value – cell\_col\_avg) / cell\_col\_stdev.**   
This gave us the option use the same scale for all columns while keeping the scaled property of the data. In addition, it helped us to easily find outliers data points.

We cleaned 65 outlier data-points. Most were obvious mistakes, like weight = 0, or ‘null’ values, while the minority was outliers like 1 person who is 82 years old in the company.

Now, using python code we extracted only the features that doesn’t influenced directly by the life routine of the person:

Table

Description automatically generated

We also decided to not use the height and weight features, since the BMI includes both in much more informative way.

**The Model:**

To answer our main question, we needed to split the data into groups that reflects the health situation of each in the best way. We wanted every group to have its own uniqueness and inside every group the data points should be as similar as possible to each other.

Since there are plenty of options to divide the data to groups with the features mentioned above, we decided to use the ML algorithm K-means, with the ‘Euclidean’ distance metric to evaluate the silhouettes of the clusters.

While approaching to build the K-means model we had 2 main questions:

1. Should we separate the man from the women in the model for better results?
2. What is the right K to choose?

So, for each K (in range 2-10) we build 3 models, one with Male only, second with Female only, and third All together. And wanted to see if there are big differences between them and that is the K where the silhouettes are the highest.

Code:

Text

Description automatically generated

Results:

Chart, line chart

Description automatically generated

The similarities of the graphs suggested that there is no benefit separating male and female from each other while using the model. In addition, we found that K=5 is best K of all.

So, we built the model for K=5 and looked at the uniqueness of each cluster:

Text

Description automatically generated

Text

Description automatically generated with medium confidence

As result we got 3 clusters that are defined by  
one main feature, and 2 more groups that  
doesn’t have big distinguish feature.

The 3 clusters that have a distinguish feature  
are: low US score, low ECG score,  
 and having family hist.

The other 2 clusters are over-all healthy people  
that are not in risk group.

At the next stage, before analyzing each of the clusters separately, we wanted to understand what the distribution of the people to the clusters is.

Chart, histogram

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

We found out that the 2 groups that doesn’t have distinct feature are the majority (80%) of the people in the company, while the family hist group is 10%, low US score is 5%, and low ECG score is 5%.

So, since most of the people are in the clusters that doesn’t have distinct feature, we choose to focus on the other 3 risk groups and explore them.

To decide what habit does the business should work on to improve his workers health the most, we now explored each cluster separately.

We did it in few steps:

1. Creating new column at the data frame with the K-means label for each raw.

Graphical user interface, application, table

Description automatically generated

1. Creating unique data frame for each group:

Graphical user interface, text, application, email

Description automatically generated

1. Display the relevant information statistics about each group, while focusing on the changeable features: smoking, work stress level, and exercise level.

Table

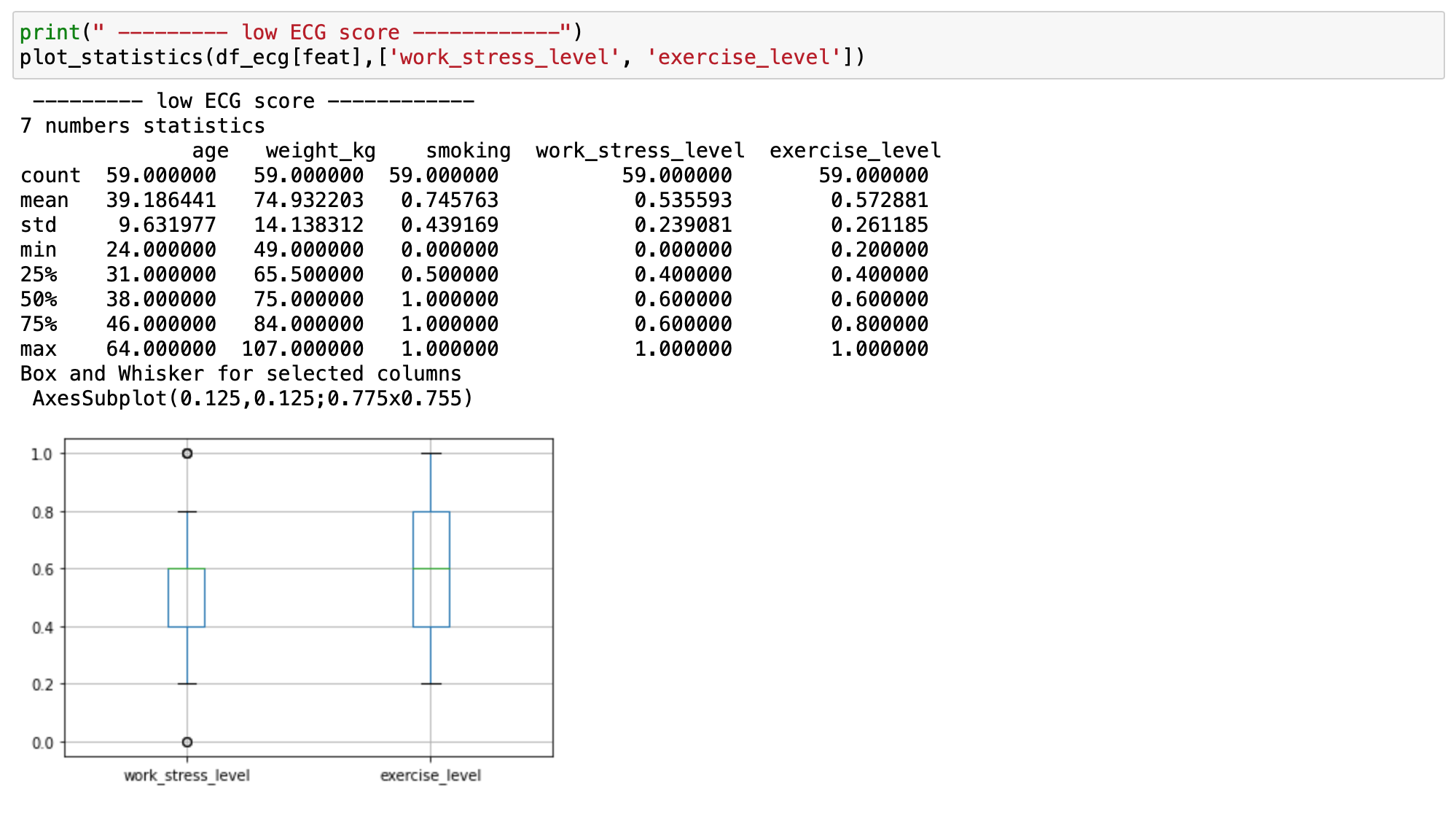
Description automatically generated

Table

Description automatically generated

Table

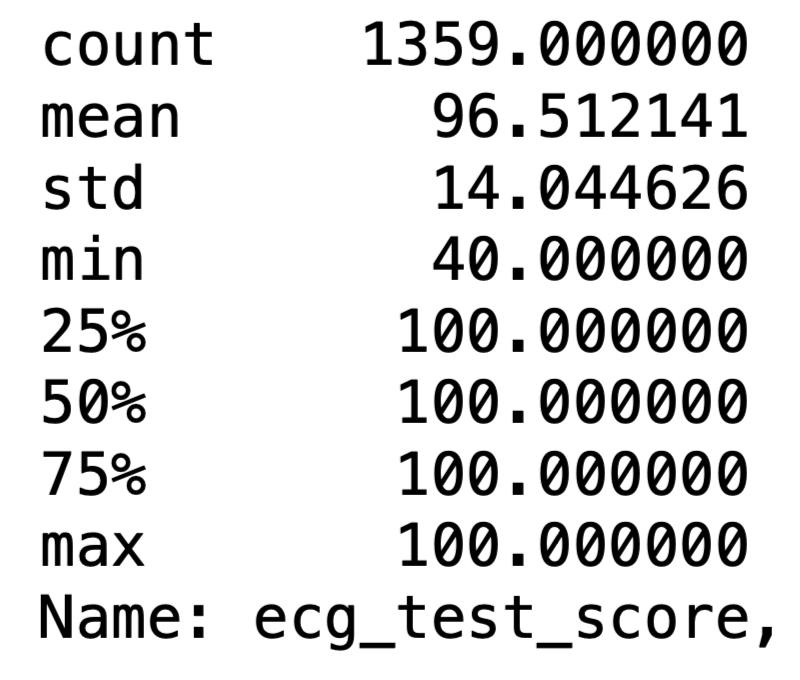
Description automatically generated



From the result we could see that the 3-risk group are exercising similar amount as the people in the ‘others’ (healthy) group and have the same amount of work stress as well. But, at the same time we saw that 20% of the people with family history and 25.5% of the people with the low ECG are smokers, which is way over the avg. of 13.5% at the healthy group!  
  
Therefore, since smoking have huge effect on health, especially on those that are already in risk group we suggest this company to try reduce smoking at the company, mainly in the mentioned risk groups.

Appendix:

1.



3.

Text

Description automatically generated

2.

Text

Description automatically generated

1.

