****

**Table Of Content**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Introduction | | |  | | | | | | 2 |
| Business S.M.A.R.T | | | | | | |  | | 3 |
| Data Exploration | | | | | |  | | | 6 |
| Data Preparation | | | | |  | | | | 8 |
| The Model | | | |  | | | | | 11 |
| Results |  | | | | | | | | 16 |
| Evaluation | |  | | | | | | | 17 |
| Bibliography & Appendix | | | | | | | |  | 20 |

**Introduction**

This report demonstrates a new, insightful tool that we believe any business would love to get, and may act upon it immediately:

Finding the “QUICK WIN” - An Easy, fast and economical solution to implement. A change that will have “immediate”, visible impact.

After data preparation and normalization, we created a model using K-means to detect “risk groups” and then found the most significant changeable criteria within these groups using unique data frames, answering the question:

**Which MINOR change in the company’s culture will achieve MEANINGFUL impact**?

In the final section of the report, we suggest an even more thorough and accurate way of measuring worker’s habits in order to detect more precise vectors of improvement, AND to be able to authenticate improvement throughout periods of time.

**Business S.M.A.R.T**

PreSee’s platform allows businesses to carefully evaluate their workers health and foresee potential risks. The data diagnostics allow a broad picture worker’s health. Taking the CRISP approach, Our goal is to illuminate an area that taking a minor action in improving the company’s lifestyle in that area will assure a real, significant gain. Whether a company wants to take care of worker’s health to the long run and run continues health check-ups, or just wants to do a one-time diagnostic and act upon it, surely the first step towards this improvement could be the “QUICK WIN” criteria.

Furthermore, as creating meaningful health changes is surely difficult, PreSee’s services would be more attractive for a business owner if they also provide a specific domain that is predicted to bring ACHIEVABLE IMPACT.

**Specific:** We want to find a SPECIFIC detail - what is the changeable habit criteria that a MINOR improvement in it will create an meaningful IMPACT?

More precisely, we are looking for the CHANGEABLE, unhealthy habit - smoking \ lack of exercise \ overweight \ work-stress \ blood-pressure - that is the most common within the company’s risk groups. These groups are defined according to results of K-means division, and are supposed to be unique according to an UNCHANGEABLE factor such as family disease history, ECG abnormalities, etc..

**Measurable:** In order to measure our success, we want to make sure our focus is on a significant group of people, and not mere individuals. Therefore, we set the total size of the focused risk group\groups (sum of risk groups) a minimum bar of 10% of all workers. Then, within these groups, we will focus on the changeable habit with highest total rates.

Furthermore, in our last section of this report we suggest a more measurable way of collecting information, setting goals and measuring improvement. The full explanation will be in “EVALUATION” section.

**Achievable:** PreSee’s diagnostics provide a wide range of data concerning the detection of high chances of heart disease and other health issues. And of course, it is collected responsibly and properly by professionals, therefore it is reliable. This allows us to identify actual risk groups with meaningful health-related features.

Realistically, It is unlikely that no risk-groups will be found in our model. Of Course, we do hope for the best health for our population and for workers of businesses throughout the globe, but heart diseases are the top 2nd reason of death in adult population in Israel1 and even more specifically, in the year 2021 there has been a rise of 15% in cardiac events (most likely due to COVID-19 IMPACT). 48% are overweight2 20% smoke.

Finding groups will be achievable, and finding an unhealthy habit within these groups is likely to be the same. Our suggested solution is itself a tool that is meant to illuminate an ACHIEVABLE IMPACT.

**Relevant:** In “Achievable” section above are detailed some of the worrying statistics of the general population. BUT our model is meant to focus on risk groups and suggest a vector to improve that will likely help the GENERAL health of all workers, yet is SURE to affect the risk groups. Why are these risk groups relevant to focus on? Because family history of heart diseases suggests a DOUBLED risk of occurrence of cardiovascular events3 According to an [NCBI study](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111047/)4, There is a strong correlation between ECG abnormalities and future sudden cardiac deaths. These categories that we define as “unchangeable”, highlight the groups within the population that are the most prone to serious health issues. It is only natural to prioritize solutions that will affect them the most.

Even in a wide-scale change of health-culture within one of PreSee’s business customers, the ability to simultaneously work on specific risk-groups and risk-groups-related factors is an important feature.

Furthermore, COVID-19 has a higher chance of deadly impact on different risk-groups, and as the data gathers on the virus’s different variants, taking care of these groups should definitely be a priority.

**Time Bound:** Our model is based on ONE-TIME tests of all workers of the company, performed at a specific time. Therefore, we do not need a constant stream of information nor we do not rely on future data.

Furthermore, our model can integrate data to achieve immediate results, as long as all of a company’s workers stats are reported properly. In our last section of this report we suggest a more measurable way of collecting information about worker’s habits (such as defining smoking on a scale of 1-5 levels) that can allow to precisely measure what would be the most significant criteria, AND how it changes the risk-groups OVERALL SCORE (and the general overall score). Implementing this can allow a business  to SET REALISTIC GOALS for NEXT YEAR’S TESTS.

This allows to intelligently measure effectiveness of decisions made by business owners on their company’s health-culture, to measure YEAR-TO-YEAR PROGRESS, and to investigate the effects on risk groups simultaneously and separately form the general group.

**Date Exploration**

We started by exploring the data for getting better understanding of the information we got.

The dataset contains 1359 records with 23 fields for each data point.

* 494 of the people are women and 866 are men.
* 194 are smoking while 1165 are not.
* 205 of them reported levels of 4, 5 of work stress, and 540 reported levels of 2 or less.
* 424 reported exercise levels of 4, 5, and 579 reported exercise levels of 2 or less.

We wanted to learn about the distribution of our data in few features:

Age distribution:

Chart, histogram

Description automatically generated

Height distribution:

Chart, histogram

Description automatically generated

Overall score distribution:

Chart, histogram

Description automatically generated

Weight distribution:

**Chart, histogram

Description automatically generated**

In addition, we created a heat-map in order tp check the corolation between different segments of our data.

Treemap chart

Description automatically generated with medium confidence  
  
We found some intresting connections such as -

* Ecg score to overall score
* Weight and BP
* BP and age

In addition, the heatmap suggests that there isn’t strong linear connection between the different tests that PreSee does, which implies that all of them giving new information – therefore – important.

Furthermore, the map helped us to better understand the affects of each feature on the overall score.

**Data Preparation**

After the basic data exploration, we realized that we need to change our data to be NUMERIC, since our model can't 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 has a different scale, we NORMALIZED 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 when evaluating chance of heart failure5. But the technical difference between having it or not - is only 1 unit!

However, the ‘height’ column, which is not a significant indicator concerning heart disease, has a range of values between 150 – 195, meaning differences can be up to 45 units!

At first, we used Excel formulas that create a binary form for 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's scale) then we decided 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)6 add 0, else 1.  
  Excel formula : IF(bp\_systolic > 125, 0, 1)
* ‘bp\_diastolic\_bit’ = If the diastolic bp was too low (< 80) 6 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 K-means algorithm we used didn't provide useful results. The algorithm kept splitting the data to similar sized groups in a linear way as long as we kept increasing the K. In addition to that, there wasn’t a meaningful distinction between groups.

Here are the results from this run:

Chart, line chart

Description automatically generatedTable

Description automatically generatedChart, line chart

Description automatically generated

So, instead of using the binary-formed information, we used a normalization formula:

**(cell\_value – cell\_col\_avg) / cell\_col\_stdev**

This gave us the option to use the same scale for all columns while keeping the scaled property of the data. In addition, it helped us find outliers data points easily.

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, who is significantly older than any other worker.

Now, using python code we extracted only the features that aren’t influenced directly by the life routine of the person, meaning they are "UNCHANGEABLE" factors:

Table

Description automatically generated

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

**The Model**

To answer our main question, we needed to split the data into distinctive risk-groups. 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 the task of building the K-means model, we had 2 main questions:

1. Should we separate men and women?
2. What is the right 'K' division to choose?

So, for each K (in range 2-10) we built 3 models: one with male only, second with female only, and third all together. Then we checked which model is best.

Code:

Text

Description automatically generated

Results:

Chart, line chart

Description automatically generated

The similarities of the graphs suggested that there is no real benefit in separating male from female. Furthermore, we found that K=5 is best K of all.

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

We received 3 clusters that are defined by one main feature, and 2 more groups that don't have any uniqueness.   
The 3 clusters that have a distinguishing feature are: low US score, low ECG score, and family hist.

The 2 other clusters are over-all healthy people that are not in our defined risk groups.

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

Chart, histogram

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

We noticed that the 2 non-distinctive groups are the majority (80%) 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 don’t have a distinctive feature, we chose to focus on the other 3 risk groups and explore them.

In order to decide which habit we should focus on, we now explored each cluster separately.

We did it in few steps:

1. Creating a 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: weight, smoking, work stress level, and exercise level.

Table

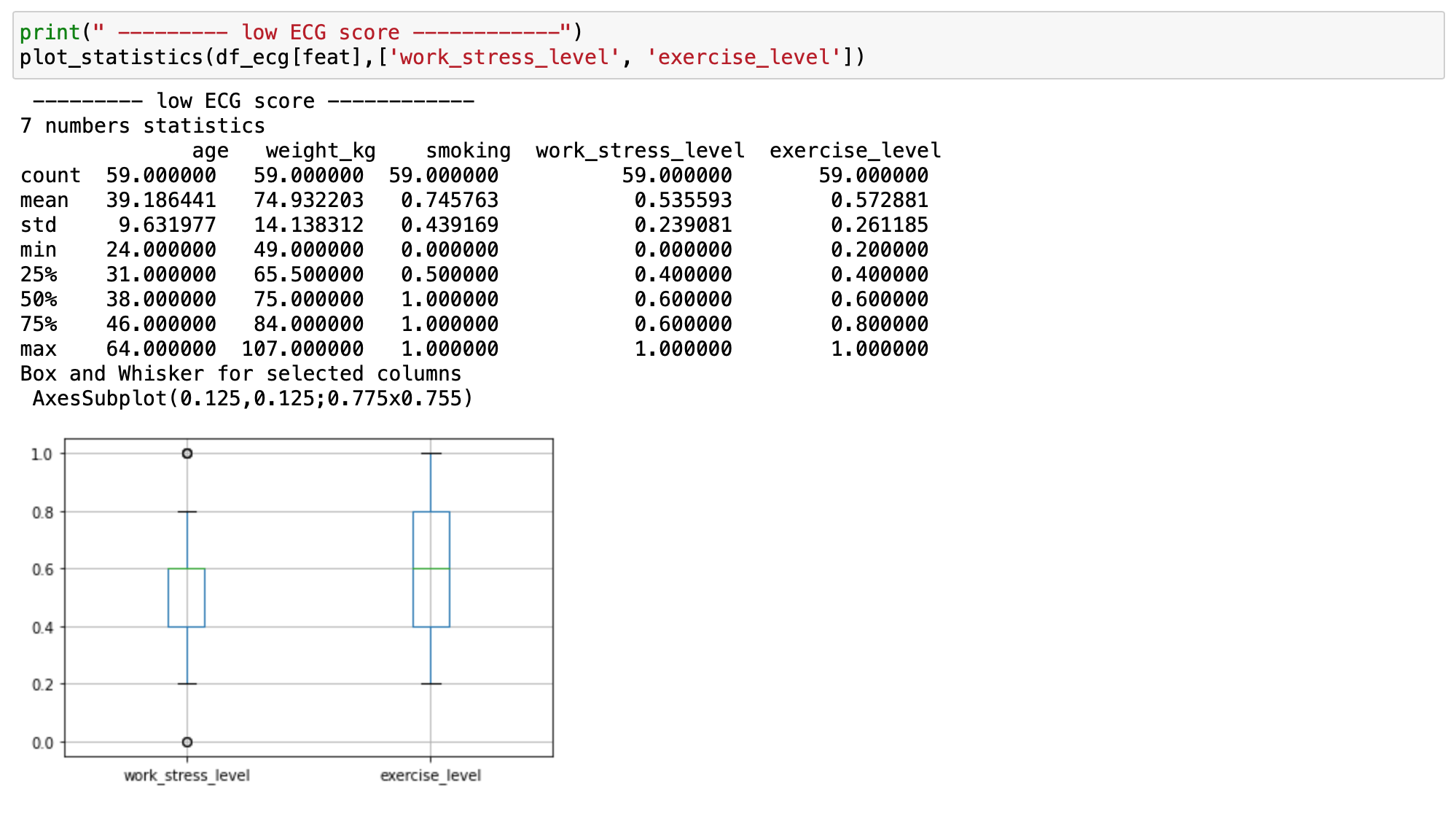
Description automatically generated

Table

Description automatically generated

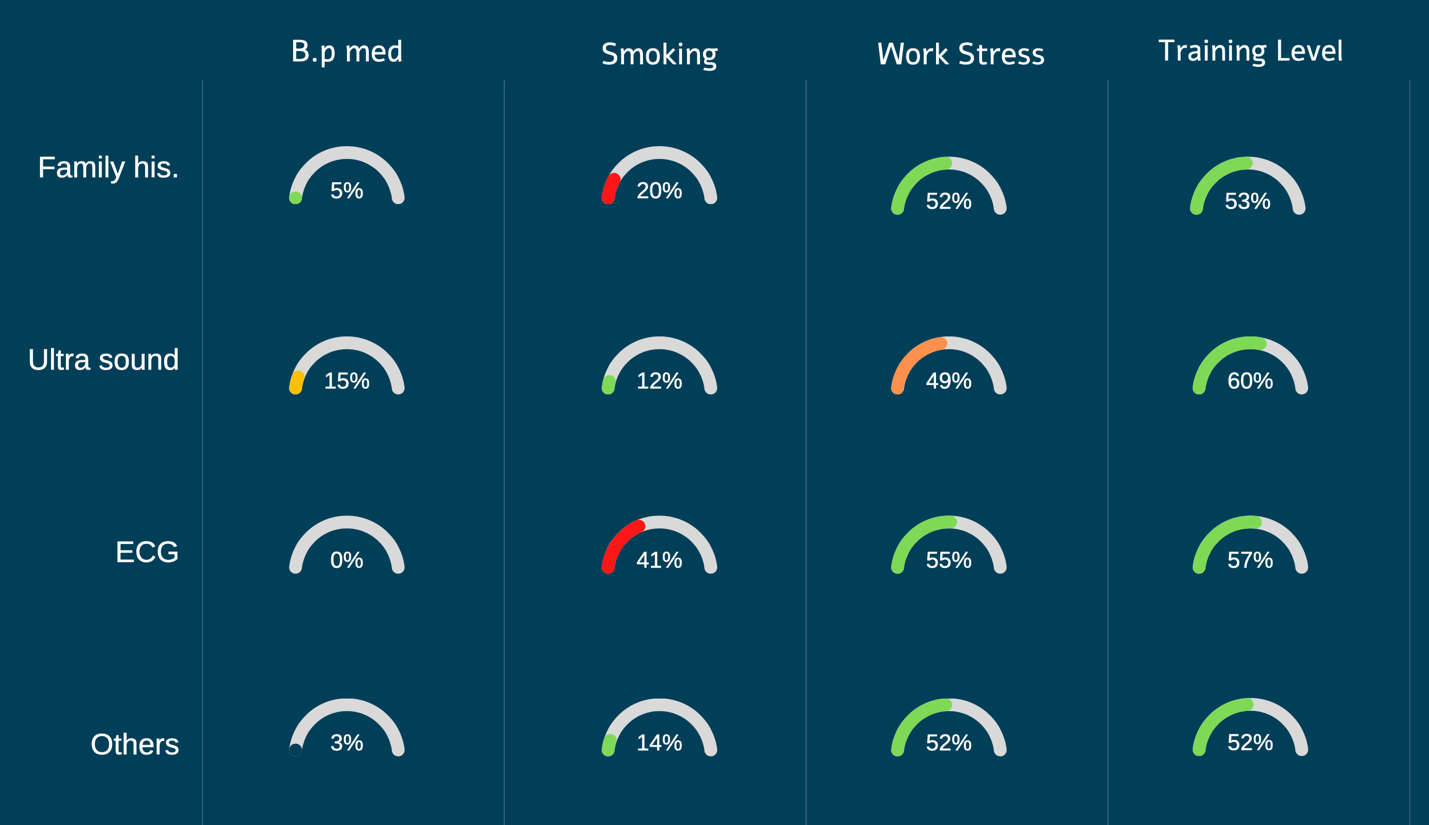
Table

Description automatically generated



**Results**

The results were 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 average of 13.5% of the general group!



**Therefore, since smoking has huge effect on health, especially on those that are already in risk groups, we suggest this company to try to focus on reducing smoking to achieve a quick, meaningful impact on risk-groups while also improving the general health of all the company's workers.**

**Evaluation**

Exploring our data, we found the best ‘K’-division for our K-Means model using male and Female together (as well as female and male separately) is **K=5**:

Chart, line chart

Description automatically generated

We notice that for K=5 we get the highest silhouette - 0.30.  
  
0.3 isn't a very high silhouet, so the clusters aren't necessarily well defined. Therefore, we weren't surprised when we saw that for every run of the algorithm we noticed a slight difference in the size of the groups.

Despite the differences, there weren't any changes in the main features that characterized the groups. This implies that our intentions of improving the “unchangeable” factors are indeed relevant and accurate.

**The Extra Feature:**

In addition to our main model, we wanted to measure the impact that our suggestions will have on the health of the company’s employees in a numeric way. In other words, to accurately measure what factor, if improving it slightly will increase the average overall score of the risk group/groups.

In order to perform this, we suggest to iterate through every unhealthy habit, and “improve” it’s value for every person in the risk group, by only 1 “TICK” (minor change):

* Training -> add ⅕ or leave ‘1’.
* Work-stress ->  reduce  ⅕ or leave ‘0’.
* Weight -> reduce by 7% - if high BMI. Else - no change required.
* Smoking  - reduce 1/5 level or leave ‘0’.

After improving each habit, we save the new AVERAGE OVERALL SCORE of the risk-groups. Before improving the next unhealthy habit, we restore the original value.

The feature that the 1-TICK change resulted in the best overall score will be the main feature that we would suggest to focus on in order to achieve a quick, measurable impact.   
  
In order to be able perform this, we suggest that Pre’See data collection will reflect these fractals. Practically, only the ‘smoking’ criteria needs modification.

Currently, PreSee decided to ask the patients whether they are smoking or not, therefore we only got a binary answer. We suggest that for this model they will start collecting the smoking habits of people in a more precise way.  
This will give us the opportunity to use the “TICK” method on this feature as well. We realize that “stop smoking” is not an easily achievable task, yet reducing smoking habit by 1 TICK is much more practical. An example for a smoking-scale:

Non smoking = 5

1-2 cigarette per day = 4

3-5 cigarette per day = 3

6-10 cigarette per day = 2

10+ cigarette per day = 1

This will create our model even more **measurable,** and we’ll know exactly how much our suggestion will improve the overall score of the company’s employees. It makes it possible to intelligently measure the effectiveness of business owners' decisions about their company's health culture, to measure year-by-year progress, and to study the effects on risk groups simultaneously and separately from the general group.

\*Also, We need to be careful about data bias and pay attention to how the question is asked in order to get a credible answer.

**Bibliography**

1. <https://publichealth.doctorsonly.co.il/2018/09/148469/>
2. <https://news.walla.co.il/item/3432891>
3. <https://www.ucihealth.org/blog/2017/02/family-history-heart-attacks>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6111047/>
5. <https://www.jstor.org/stable/3703625>
6. <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>

**Appendix**

3.

Text

Description automatically generated

2.

Text

Description automatically generated

1.

