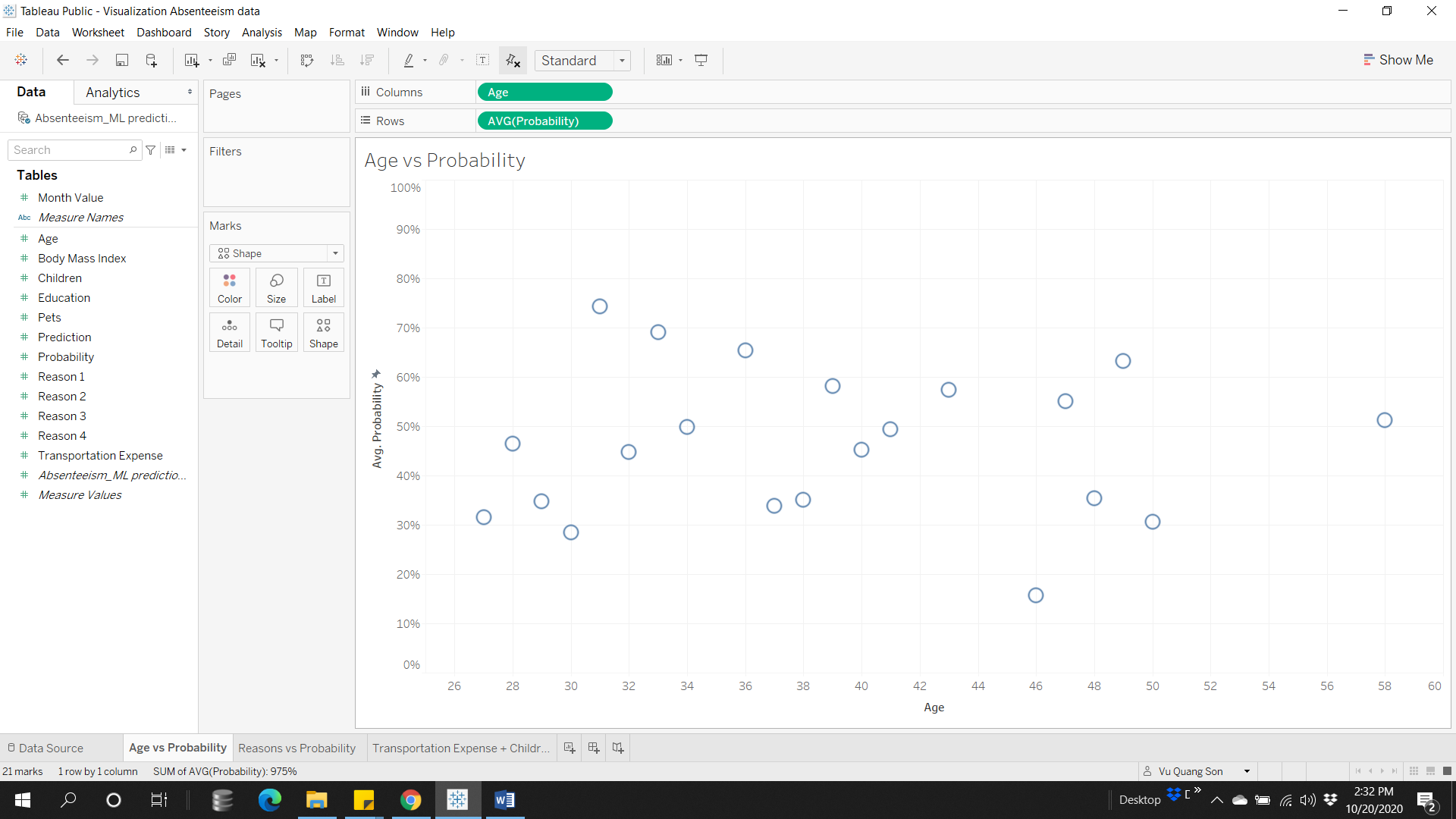
After I had pre-processed raw data and after obtaining clean data I used the Logistic Regression ML model (the model is given) to predict the probability of being absent for more than three hours of 700 employees. Then I transferred all of those data into a table in MySQL, next I exported all of those data to a CSV file. Finally, I used the predicted data to see if there is any correlation between factors like Age, Reasons for being absent, Transportation expense + the number of children and the probability/ Avg.probability of absenteeism for more than three hours on a scatter plot. Now let’s interpret the insights from my finding.

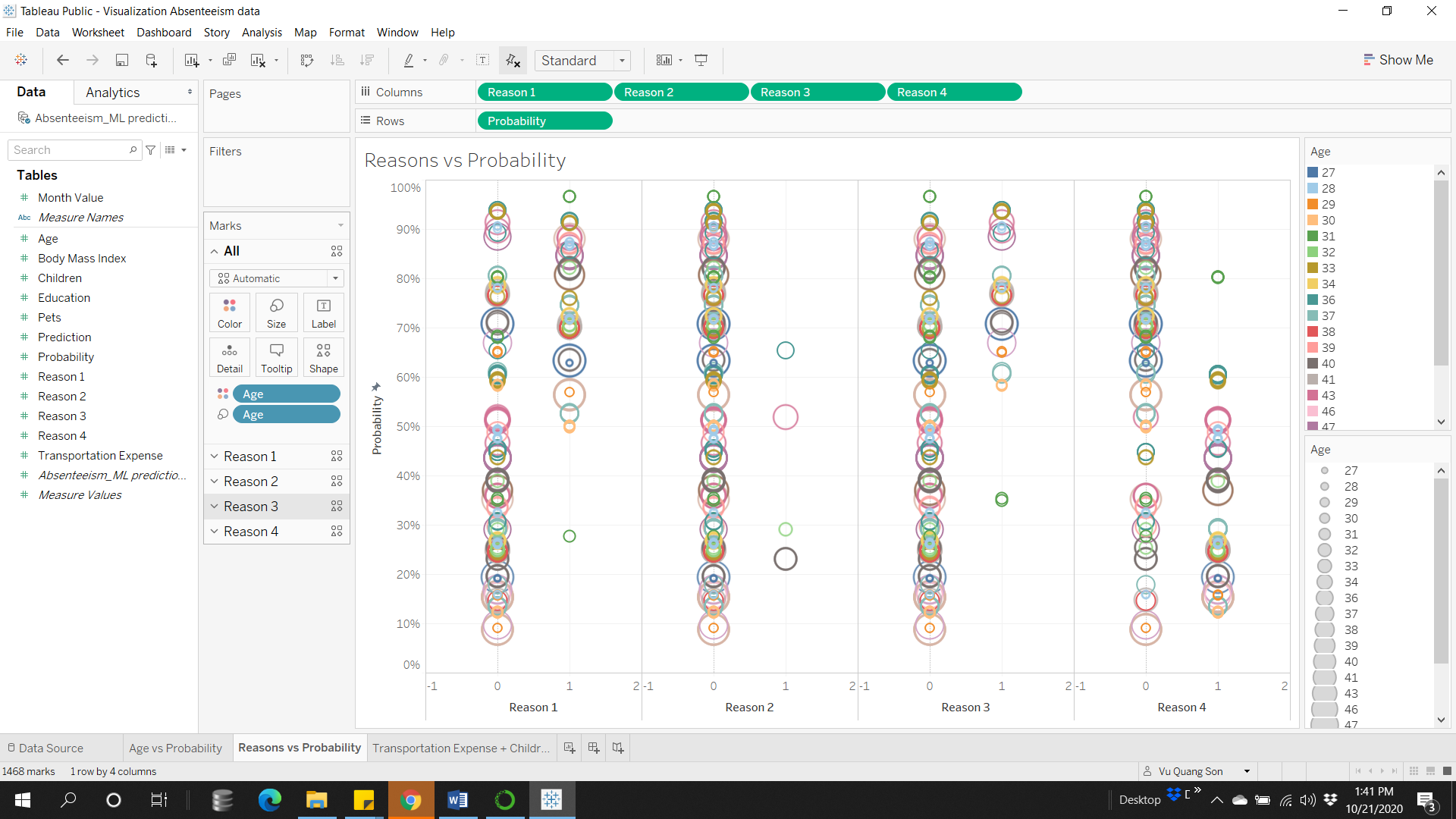
**Link of three worksheets**: <https://public.tableau.com/profile/vu.quang.son#!/vizhome/VisualizationAbsenteeismdata/ReasonsvsProbability?publish=yes>

**I) WORKSHEET 1:**



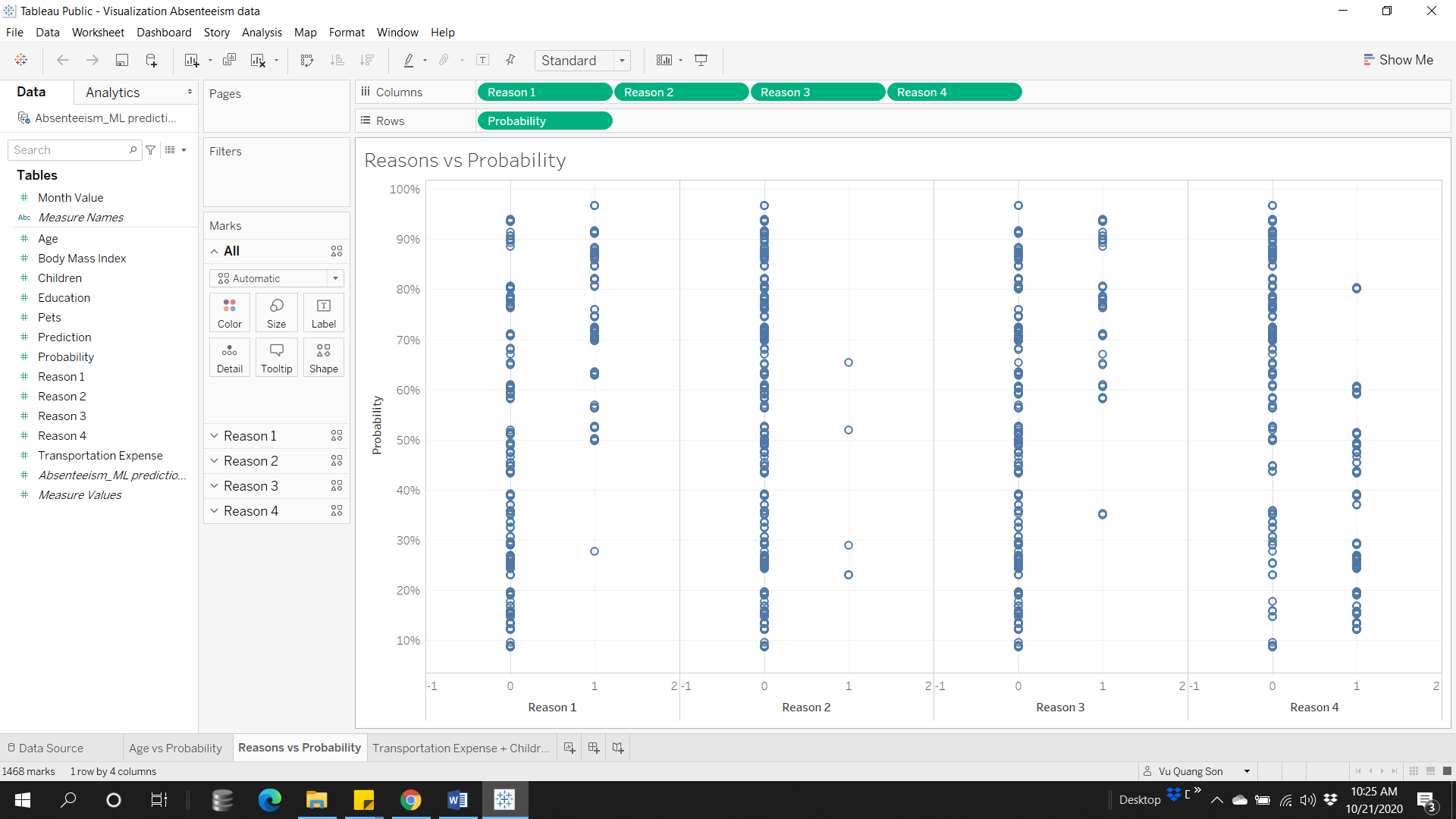
**Note**: *see the link contains three worksheets to know the average probability corresponding to each age*

In general, there’s a fragile correlation between the ‘Age’ variable and the ‘Avg. Probability’ (Average Probability) variable, the **blue line** represents the linear regression line approximate based on the distribution of observations, and the **red line** shows observations above and below the 50% probability. Firstly, we can see that people with the age below 30 are less likely to be absent for more than three hours, with the average probability of each age from 26 to 30 is 35.5%. Secondly, between the age of 50 and 60, a group of older employees is less likely to be absent from work, this can be because at this age if they don’t have any health condition when they were younger, they might live healthy for the next 10, 20 or 30 years or if they have children, their children have grown up so they don’t have to look after their children anymore. Thirdly, we can see the age range that people will most likely have absenteeism is from over 30 to below 60 years old, with an approximately average probability of 46%. Especially, from over 30 to below 36, the employees in this range have approximately 61% of being absent for more than three hours. Finally that the average probability shown here is a chance for a person to be absent for more than three hours, a low average probability just means they will be less like to be absent for more than three hours, it doesn’t mean they are less likely to be away for one or two hours. In conclusion, we have found the range that people are most likely to be absent, these people being absent may due to they have diseases or they have dependents to take care of. To know why’s that, we can contrast this information with factors such as reasons for being absent, the number of children, BMI, or transportation expenses, by adding different fields to a chart.

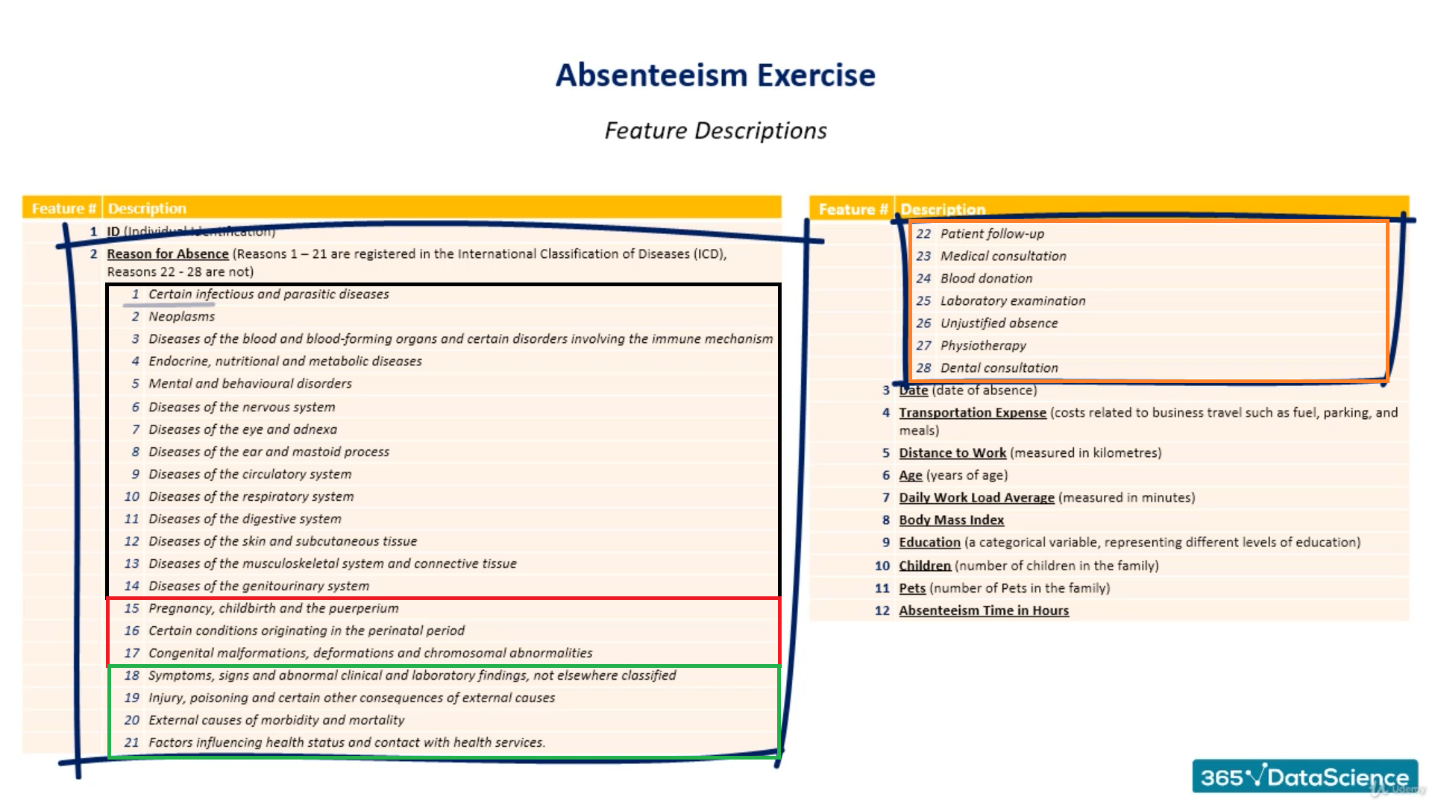


**Note**: *a modified version to show a relationship between four groups of absent reasons and the probability they will be absent for more than three hours and each observation (an employee) is assigned to their given age. You can interact with the chart to find out more information like an employee with the age of 31 has an average 74% of being absent, what groups of reasons they belong to.*

**II) WORKSHEET 2:**

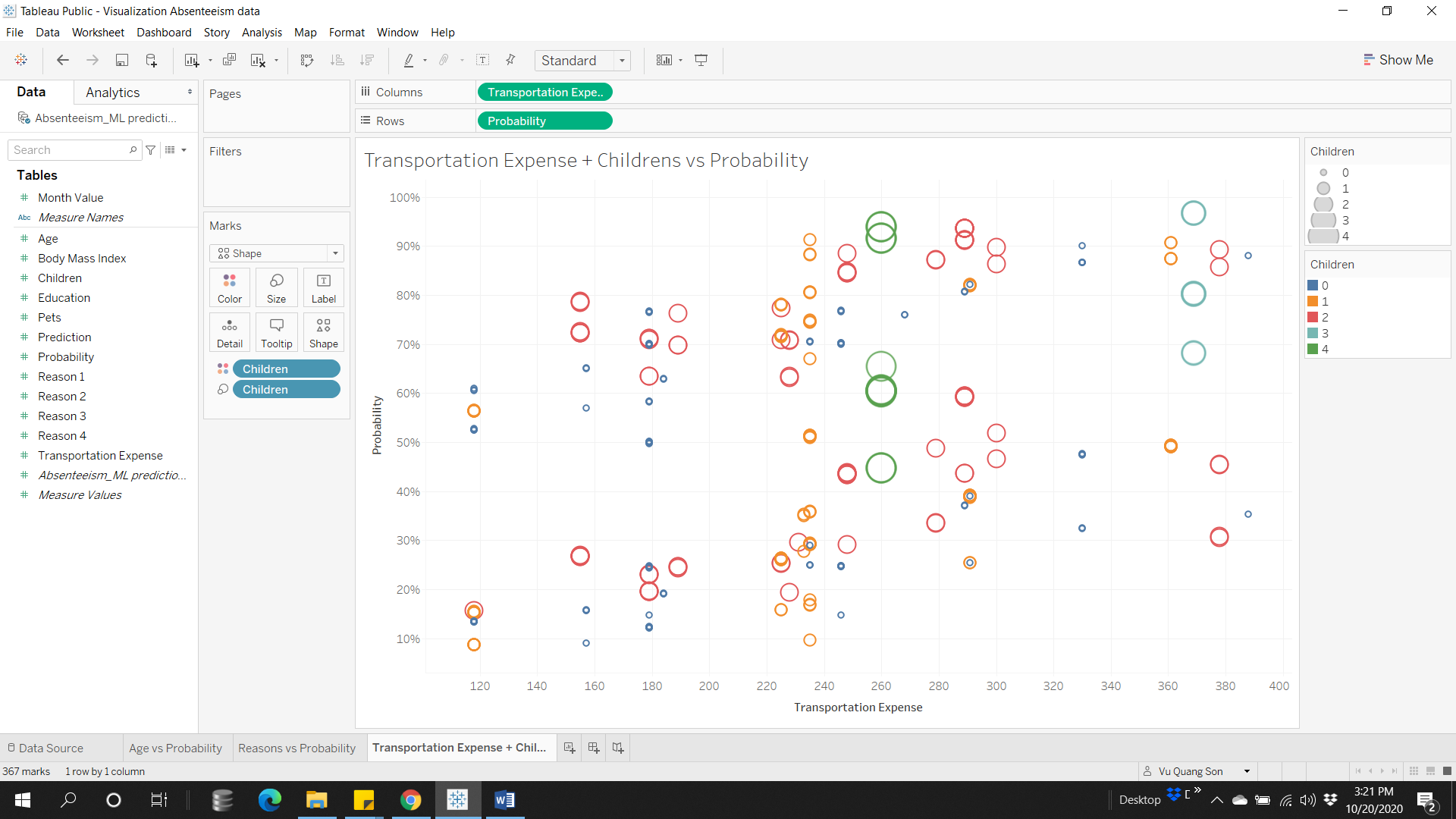


**Note**: *see the link contains three worksheets to know the probability corresponding to each reason. 0 means an employee didn’t have a disease from a particular group and 1 means they have.*

**

***The table above is a list of diseases that were being reported by employees.***

A quick overview, I separated the list into four groups, group one includes the first **14 diseases**, group two will contain **15 to 17 reasons**, group three includes diseases numbered **18, 19, 20, 21,** and finally, **group four** has the **rest**. From the graph we obtained, we can see that only four employees in the total of 700 employees have reasons from group two, those reasons relating to pregnancy, we can see that the people being absent for those reasons are insignificant, so we don’t need to worry too much about employees in that group. Secondly, we can see that employees who have an excessive probability of being absent for more than three hours belong to group number one and group number three, they have serious diseases related to organs, circulatory, respiratory system or diseases that are not elsewhere classified, hence why they have a higher chance of being absent. Finally, group number four also has a large number of employees having it, but none of them could be serious enough to require a person to be absent excessively, the reason is that if we look up in the reasons for absence table we can see that the diseases in group four are light diseases such as dental consultation or blood donation. In conclusion, the table above will tell us what to expect if an employee who is likely to have or is having a certain disease belongs to a specific group above what is his estimated probability that he will be absent for more than three hours? You can add the ‘Age’ field to obtain more specific information.

**III) WORKSHEET 3:**   


**Note**: *see the link contains three worksheets to interact with the chart*

Firstly, when we interact with the graph, we will see that people with no or only one child don’t exhibit an excessive probability of being absent, the chance of being absent does increase in association with the Transportation Expense, but the majority of employees in this group don’t spend more than 250$ for Transportation Expenses. Secondly, the probability of being absent is high for half of the employees who have above two children, no matter their Transportation Expense is high or low. Thirdly, employees who have three children tend to have high transportation expenses and there’s an exception here people with four children have exactly the same Transportation expenses, maybe they live in the same neighborhood or there might be other reasons that they don’t want to live too close or too far from where they work. In conclusion, there aren’t many employees who have more than two children and there seems to be a correlation between the number of children and the Transportation expenses of an employee, except for employees who have two children.