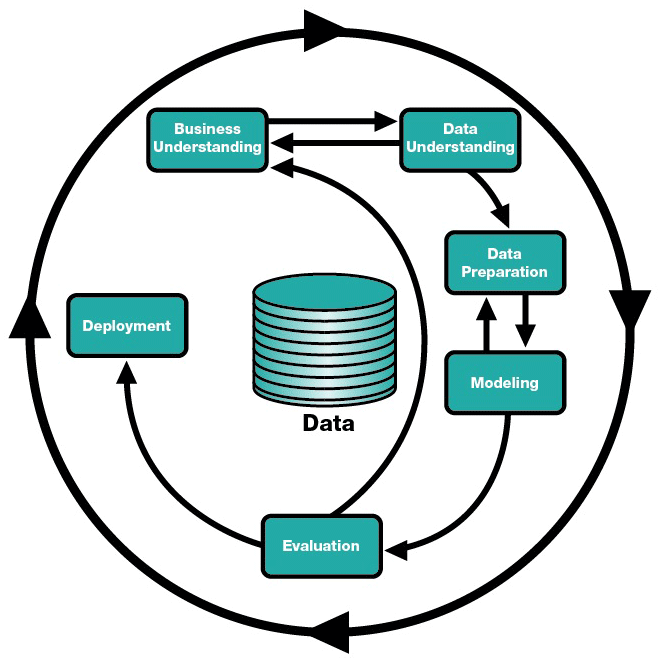
**Predicting the Factors that Impact Mental Health in the Tech Work Place**

Nichita Jamwale, Rode Esha

**Introduction**

The aim of the project was to predict the factors that affect mental health in the tech work place, based on a survey data provided by Open Sourcing Mental Illness (OSMI) on Kaggle. SAS JMP was very vital in carrying out this analysis and for creating our models. We used the Nominal logistic regression model and Decision trees in creating our model and we also look to test the performance and accuracy of our model. The Cross-industry standard process for data mining, known as CRISP-DM, which is the most widely used analytic model was our approach for this project.



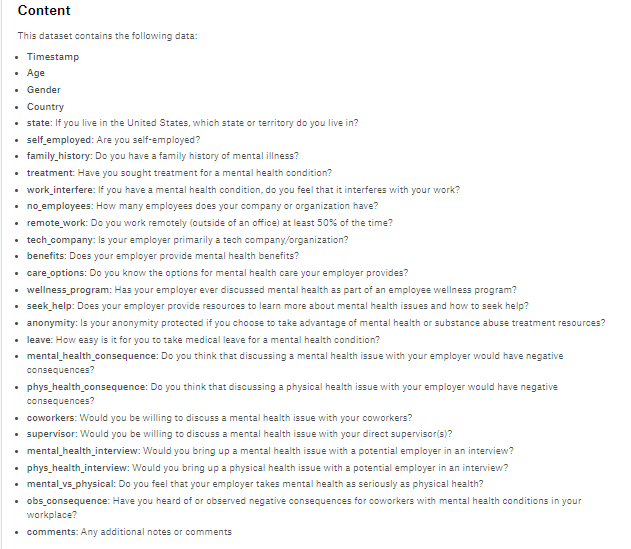
**Business understanding**

Open Sourcing Mental Illness (OSMI) is a non-profit 501c3 corporation dedicated to “raising awareness, educating, and providing resources to support mental wellness in the tech and open source communities.” What they do in support of this goal includes providing e-books on mental wellness in the workplace, hosting a forum on conversations on mental health, and holding talks at developer conferences about mental health in the community.

The interest or business value of this project is to understand the view of mental health within the tech workplace and prevalence factors that affects mental health. The Open Sourcing Mental Illness (OSMI) team of volunteers will use this data to drive their work in raising awareness and improving conditions for those with mental health disorders in the tech work place

**Data understanding**

OSMI provided a survey data on mental health in tech industry on Kaggle. This survey contains a variety of questions pertaining to the mental health of the respondents, the demographics of the respondents, and how employer views mental health in the workplace. This survey was conducted in 2016.



The dataset had about 1,259 observations and 27 variables, which consisted of 2 continuous variables and 25 categorical variables. It was imperative we explore the data and understand what each variables speaks to before we began our analysis.

**Data Preparation**

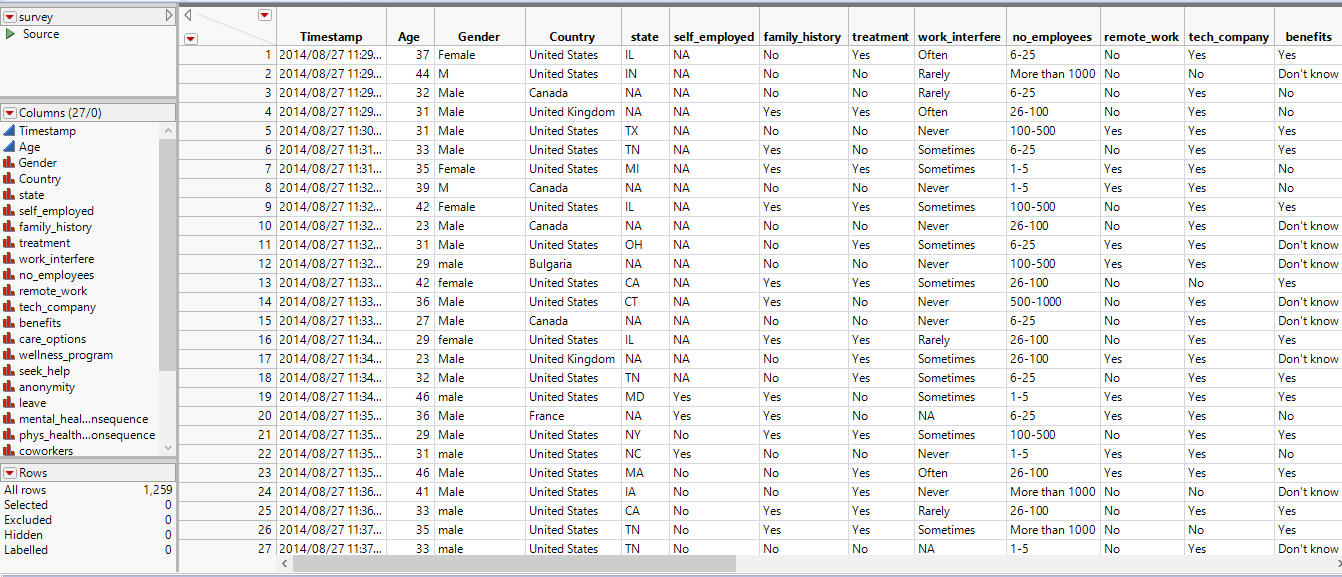
* **Tools**

SAS JMP was a great tool to manipulate the data and make statistical models. JMP was used to identify and resolve outliers, missing value and erroneous data, and was also used to test the significance of the variables included in the dataset.

* **Visualization**

We were able to visualize the data first using the Distribution to get an understanding of the dataset and to know how spread the data was and to understand the levels in each nominal variables.

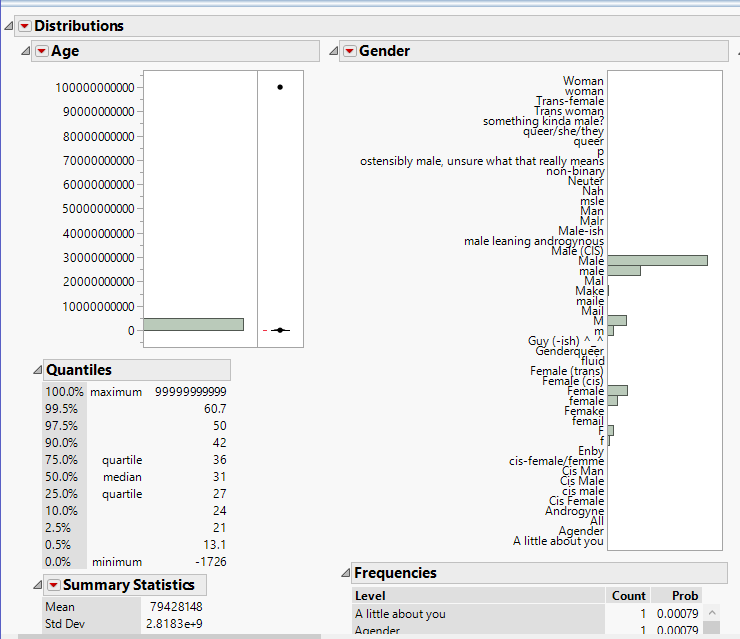
From the visualization, variables not normally distributed, containing outliers and erroneous data were identified easily for clean-up.



* **Clean up**

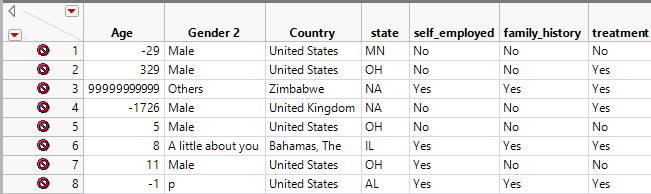
From visualizing and understanding the dataset, we were able to identify 2 variables, ‘Timestamp’ which is about the time the survey was taken all in the same day but different minutes and ‘comments’ which were series of comments made about the interview. We saw this as not important for our analysis, which were deleted from our dataset.

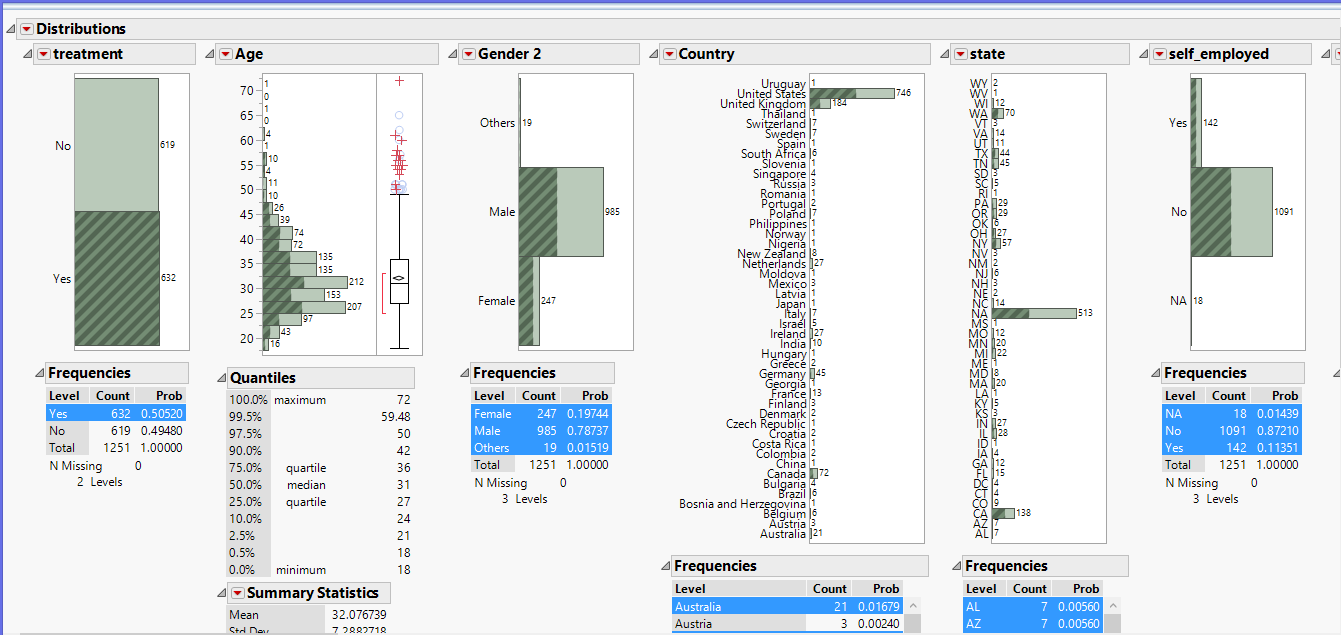
We further analyzed the 25 outstanding variables needed for our analysis and identified that Age which was the only continuous variable was highly skewed due to some outliers and Gender had erroneous data, which led to multiple Gender category. See image below



Erroneous data for Gender that resulted to about 47 categories were resolved, by regrouping to three categories (Male, Female and Others)

Also, in the Age variable 8 observations were excluded, we felt it wise to take out these variables because they were outliers with influencing our model

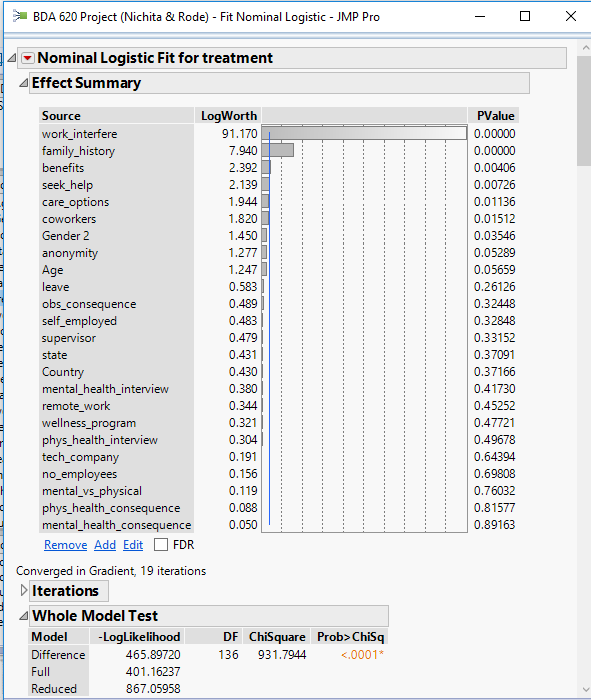




**Modeling**

* **Nominal Logistic Regression**

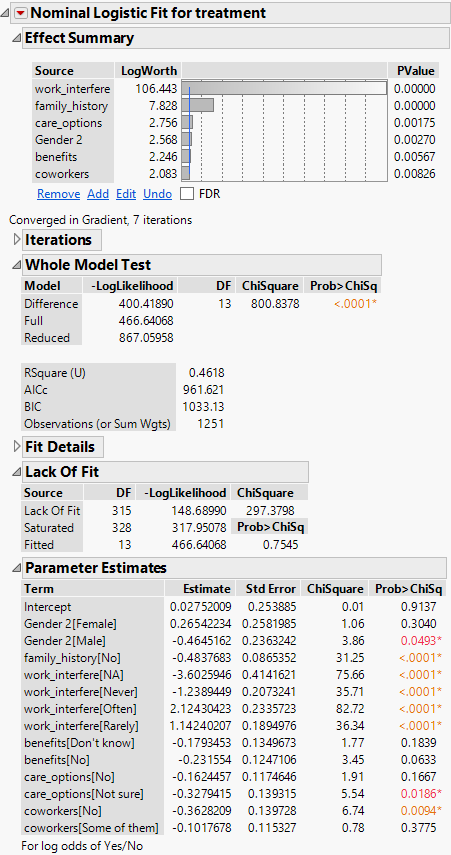
After preparing our data we went ahead to create our model, we use the Analyze > Fit model to create a Nominal Logistic Regression model by regressing 24 of our predictors on the response (Treatment).



We began eliminating each variables that were not significant for our model till we reached a significant number of 6 variables

* Work\_interfere
* Family\_history
* Care\_options
* Gender
* Benefits
* Coworkers

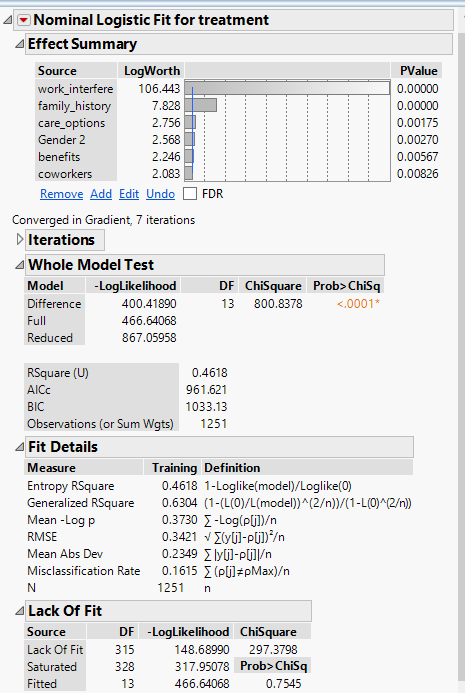
which gave us a significance model <0.05 with a –loglikelihood of 466.64 and ROC value of 0.908

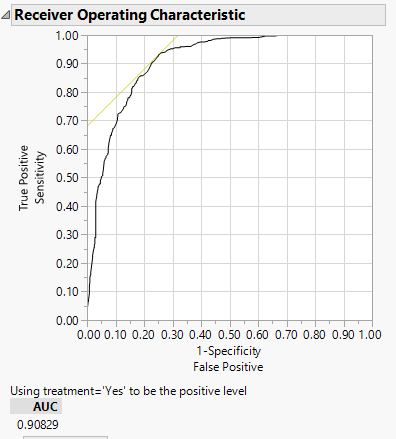


**Model testing**: Having created a model and coming up with 6 significant variables, and seeing the model was significant with a minimal -Loglikelihood value, testing the performance of our model was the next action taken, we looked at the Confusion matrix, Area Under the Curve (AUC) and Misclassification rate

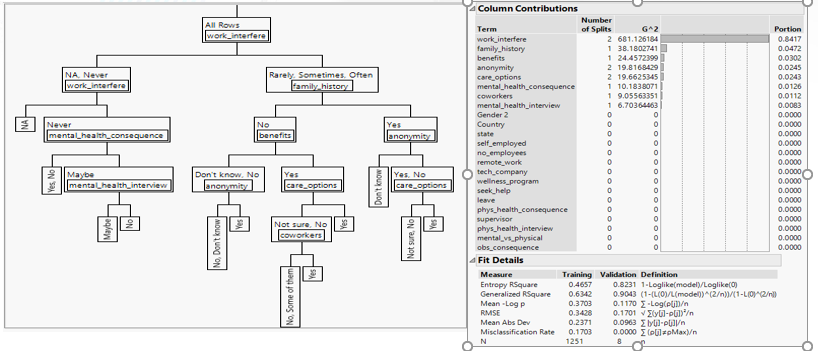
Outcome of our performance check

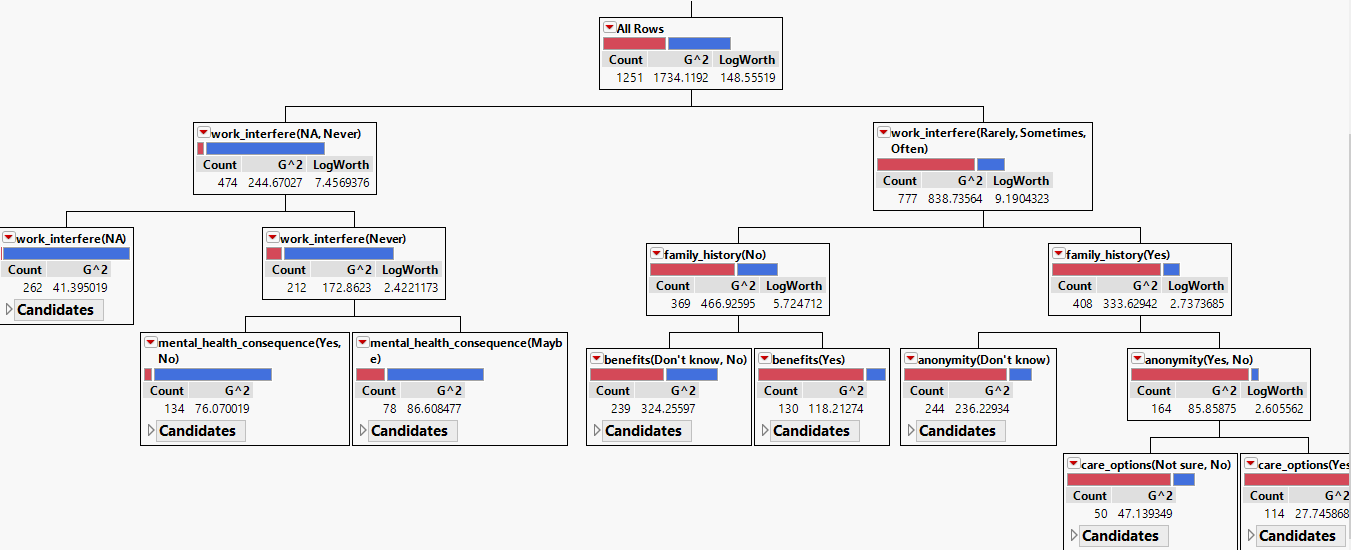
* First, we saw that the model was significant with P value less than 0.05
* This model also provided us with a –loglikelihood of 466.64, a AUC of 0.90829 and a misclassification of 0.16.
* Also using the Confusion matrix, we were able to check the accuracy of the model which was about 83.8%

****



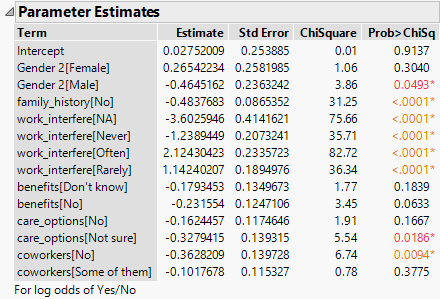
* **Decision tree (Classification tree)**



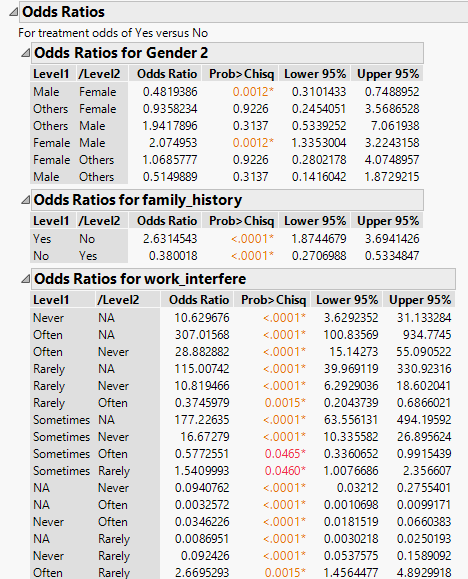


**Evaluation**

* **Findings and Results**
* Looking at the parameter estimates, which enables us to see the effect of each predictors on the Treatment variables that shows the likelihood of having a mental health if ‘Yes’. For example looking a the Gender2[Male] we can see that the estimate is negative, indicating that employees who are male have a lower impact on the Y variable than female employees.
* Also Work\_interfere[Often] and Work\_interfere[Rarely] had a 2.124 and 1.142 respectively, which shows a greater impact on the model. That is employees who said having a mental health issue will impact their work often and rarely had higher chance of having mental illness.

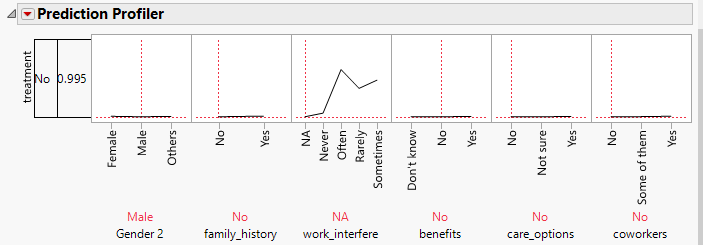
****

* Looking at the odds ratio for Gender 2 with levels (male/female) we can see that odds of a male having a mental health or sought treatment ‘Yes’ is 0.482 lower than that of a female and also looking at the Gender 2 with levels (female/male) the odds of a female to male is 2.075 times higher.

****

****

* We can better understand our model by using the prediction profiler, we were able to see how changing one factor affects the response “Treatment” and as well as affects other factors in the model and also we can see what profile can give us the highest probability of having a mental and which is seen above.
* If the employee was a female, has a family history as ‘Yes’, has work\_interfere as ‘Sometimes’, benefits ‘Yes’, care\_options ‘Yes’ and Coworkers ‘Yes’ resulted to the highest probability of 97.6% to have a mental health issue.



* If the employee was a male, has a family history as ‘No’, has work\_interfere as ‘Na’, benefits ‘No’, care\_options ‘No’ and Coworkers ‘No’ resulted to the lowest probability of 0.5% to have a mental health issue.
* However, looking at the model of the 5 significant variables, Work\_interfere, Family\_history, Gender and care\_options had greater impact on the model.

**SUMMARY**

* It could be surprising for some to see that female are more likely to have mental health issues than men. From the 247 female employees who responded to this survey, 170 which is about 68.8% reported that they have sought mental treatment while for the male employees of the 985 that responded, 447 which about 45% said they have sough mental health treatment.
* Based on the analysis we were able to see the characteristics of employees who are most likely to have a mental health issue.
* Females, with a family history of mental illness, particularly if they responded that having this issue could interfere with their work and who cared more about benefits and care options, and are willing to discuss this with a co-worker were more likely to have mental illness.
* From this findings OSMI can properly engage employees in the tech place and create more awareness especially to female employees

**Reference**

* Kaggle <https://www.kaggle.com/osmi/mental-health-in-tech-survey>
* JMP <https://www.jmp.com/en_us/home.html>
* OSMI (Open Source Mental Illness) <https://osmihelp.org/research>

After Elimination of insignificant variables we have

