Factors affect engagement and satisfaction of employees in the Federal workforce

A Capstone project:
Tonia Chu
Under the mentorship:
Dr. Marko Mitic (Data Scientist at Telenor, Belgrade, Serbia)
For the course:
Foundations of Data Science (Springboard)

I. Introduction

In 2015, more than 400,000 employees participated in the Federal Employee Viewpoint Survey (FEVS). The survey includes questions about satisfaction, leadership, and work schedules. The analysis try to identify the working status and give result on in which respect and how to improve the feelings of federal employees. From the analysis of the survey, we can get answers to many interesting questions.

Employee feedback on key performance metrics captured in the FEVS is singularly important for each agency to realize its mission, as well as maintain and enhance Federal workforce. The feedback enables each agency to develop effective strategies and tools for driving continuous improvement.

In this study, I want to solve following problems:

1. How to measure satisfaction and engagement of employees with answers of survey?
2. What are the relationships of satisfaction/engagement via different groups (gender, age, years of service, supervisor status, agency etc.)?
3. What factors affect satisfaction of employees?
4. What factors affect engagement of employees?
5. How to drive engagement of employees?

II. Purpose

Each year leaders in the Federal Government use the Federal Employee Viewpoint Survey (FEVS) as a management tool to drive change and increase employee engagement in the federal workforce. The use of that data continues to translate into better service for the American people.

The purpose of this project is to analyze the data of FEVS and identify the working status and give results on in which respect and how to improve the feelings of federal employees, measure employees' perceptions of whether, and to what extent, conditions characterizing successful organizations are present in their agencies.

The analysis try to provide information for agency leaders and managers seek to improve their human capital management. Every agency has its own unique mission and workforce, and its own set of special human capital challenges. Guided by the analysis results, agency leaders can take steps to improve how employees engage with their jobs, organizations, and missions.

III. DATASET

This project use the annual survey of the Office of Personnel Management.

Website: http://www.fedview.opm.gov/2015/EVSDATA/

The dataset is in a CSV file and include 84 questions about Work Experience, Work Unit, Agency, Satisfaction, Work/Life balance and so on. It also gives demographics information like gender, age, supervisor status, years in the Federal Government, about to leave or retire, work unit and so on.

The survey questions are designed to get feedbacks on the three main indices: Engagement Index, Global Satisfaction Index and The New IQ Index.

* Engagement Index

Employee engagement is the employee’s sense of purpose. It is evident in their display of dedication, persistence, and effort in their work or overall commitment to their organization and its mission. An agency that engages its employees ensures a work environment where each employee can reach his or her potential, while contributing to the success of the agency. Individual agency performance contributes to success for the entire Federal Government. The index is made up of three sub-factors: Leaders Lead, Supervisors, and Intrinsic Work Experience. Each sub-factor reflects a different aspect of an engaging work environment.

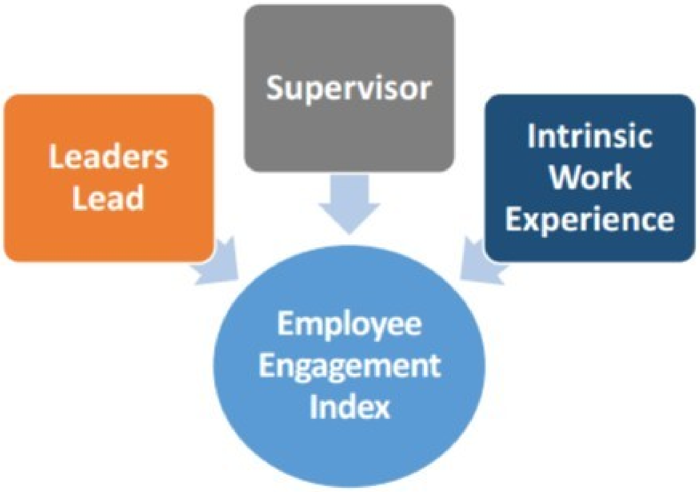


Figure 1: Sub-factors for the Employee Engagement Index

Leaders Lead reflects the employees' perceptions of the integrity of leadership, as well as leadership behaviors such as communication and workforce motivation. It is made up of the following survey items:

In my organization, senior leaders generate high levels of motivation and commitment in the workforce. (Q. 53)

My organization's senior leaders maintain high standards of honesty and integrity. (Q. 54)

Managers communicate the goals and priorities of the organization. (Q. 56)

Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor? (Q. 60)

I have a high level of respect for my organization's senior leaders. (Q. 61)

Supervisors reflects the interpersonal relationship between worker and supervisor, including trust, respect and support. It is made up of the following survey items:

Supervisors in my work unit support employee development. (Q. 47)

My supervisor listens to what I have to say. (Q. 48)

My supervisor treats me with respect. (Q. 49)

I have trust and confidence in my supervisor. (Q. 51)

Overall, how good a job do you feel is being done by your immediate supervisor? (Q. 52)

Intrinsic Work Experience reflects the employees' feelings of motivation and competency relating to their role in the workplace. It is made up of the following survey items:

I feel encouraged to come up with new and better ways of doing things. (Q. 3)

My work gives me a feeling of personal accomplishment. (Q. 4)

I know what is expected of me on the job. (Q. 6)

My talents are used well in the workplace. (Q. 11)

I know how my work relates to the agency’s goals and priorities. (Q. 12)

* Global Satisfaction Index

The Global Satisfaction Index measures employee satisfaction about four aspects related to their work: their job, their pay, their organization, and whether they would recommend their organization as a good place to work. Understanding employee satisfaction along these four dimensions can help reduce costs in the long run. Satisfied employees are more likely to stay in their jobs, reducing turnover.

The **Global Satisfaction Index** is comprised of the following survey items:

I recommend my organization as a good place to work. (Q. 40)

Considering everything, how satisfied are you with your job? (Q. 69)

Considering everything, how satisfied are you with your pay? (Q. 70)

Considering everything, how satisfied are you with your organization? (Q. 71)

* The New IQ Index

The New IQ identifies behaviors that help create an inclusive environment and is built on the concept that repetition of inclusive behaviors will create positive habits among team members and managers. Behaviors included in the New IQ can be learned, practiced, and developed. Consequently, all members of an organization can improve their inclusive intelligence. Workplace inclusion is a contributing factor to both employee engagement and organizational performance.

The New IQ is comprised of the following sub-factors and items:

Fair: Are all employees treated equitably? (Q 23, 24, 25, 37, & 38)

Open: Does management support diversity in all ways? (Q 32, 34, 45, & 55)

Cooperative: Does management encourage communication and collaboration? (Q 58 & 59)

Supportive: Do supervisors value employees? (Q 42, 46, 48, 49, & 50)

Empowering: Do employees have the resources and support needed to excel? (Q 2, 3, 11, & 30)

In this project, we’ll do some research about factors affect the satisfaction of employees and focus on employee engagement. A successful agency fosters conditions essential to an engaged workforce to ensure each employee can reach his or her potential and contribute to the success of the agency. Research shows a relationship between employee engagement and performance.

Analysis of Federal Employee Viewpoint Survey (FEVS) data shows specific factors support conditions for achieving an engaged workforce.

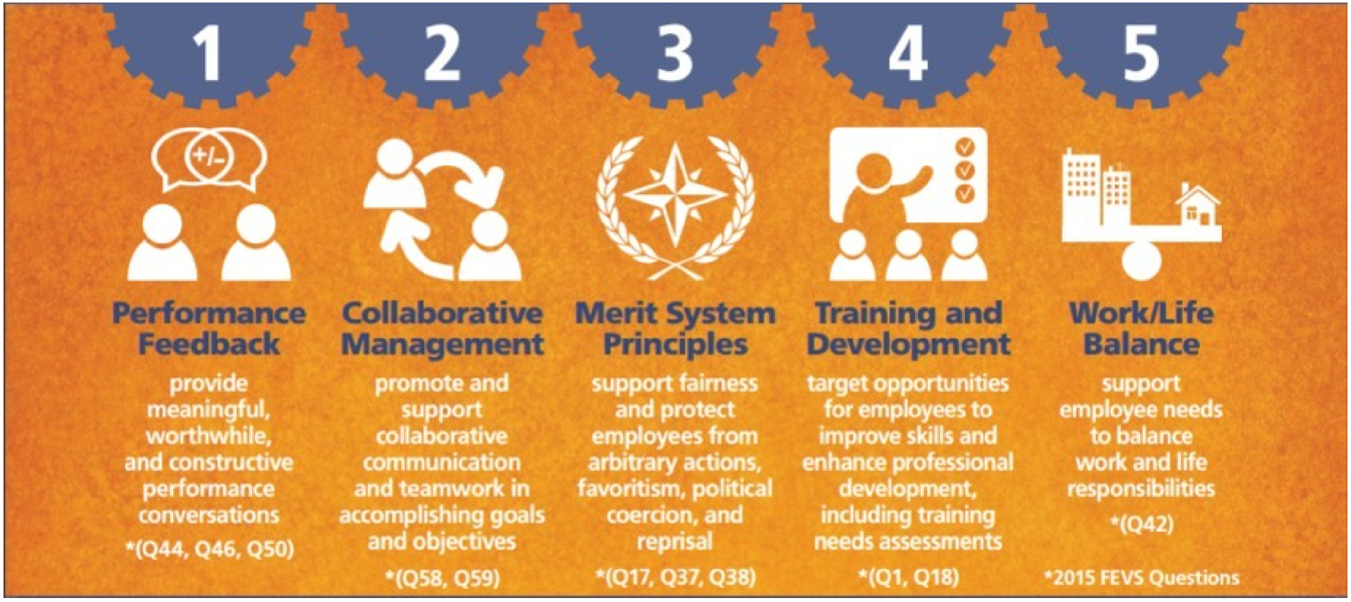

1
Performance Feedback
provide meaningful, worthwhile, and constructive performance conversations
*(Q44, Q46, Q50)
2
Collaborative Management
promote and support collaborative communication and teamwork in accomplishing goals and objectives
*(Q58, Q59)
3
Merit System Principles
support fairness and protect employees from arbitrary actions, favoritism, political coercion, and reprisal
*(Q17, Q37, Q38)
4
Training and Development
target opportunities for employees to improve skills and enhance professional development, including training needs assessments
*(Q1, Q18)
5
Work/Life Balance
support employee needs to balance work and life responsibilities
*(Q42)
*2015 FEVS Questions

Figure 2: Drivers for the Employee Engagement Index (EEI)

In this project, we'll work on the questions related to the three indices and EEI drivers.

IV. Data Wrangling

First, we load dataset from csv file, at the same time set empty string, space and "X" as NA.

```
> survey_all <- read.csv("evs2015_PRDF.csv", na.strings=c(""," ", "X"), stringsAsFactors = FALSE)
```

Then we select the useful columns based on the survey questions related with the three indices and drivers of EEI.

```
> survey <- survey_all[ , c("agency", "Q1", "Q2", "Q3", "Q4", "Q6", "Q9", "Q10", "Q11", "Q12",
"Q15", "Q16", "Q17", "Q18", "Q19", "Q22", "Q23", "Q24", "Q25", "Q30", "Q32", "Q34", "Q37", "Q38",
"Q40", "Q42", "Q44", "Q45", "Q46", "Q47", "Q48", "Q49", "Q50", "Q51", "Q52", "Q53", "Q54", "Q55",
"Q56", "Q58", "Q59", "Q60", "Q61", "Q69", "Q70",  "Q71", "DSUPER", "DSEX", "DMINORITY", "DE-
DUC", "DFEDTEN", "DRETIRE", "DDIS", "DAGEGRP", "DMIL", "DLEAVING")]
> names(survey)
 [1] "agency"    "Q1"       "Q2"       "Q3"       "Q4"
 [6] "Q6"       "Q9"       "Q10"      "Q11"      "Q12"
[11] "Q15"      "Q16"      "Q17"      "Q18"      "Q19"
[16] "Q22"      "Q23"      "Q24"      "Q25"      "Q30"
[21] "Q32"      "Q34"      "Q37"      "Q38"      "Q40"
[26] "Q42"      "Q44"      "Q45"      "Q46"      "Q47"
```

```
[31] "Q48"      "Q49"      "Q50"      "Q51"      "Q52"
[36] "Q53"      "Q54"      "Q55"      "Q56"      "Q58"
[41] "Q59"      "Q60"      "Q61"      "Q69"      "Q70"
[46] "Q71"      "DSUPER"   "DSEX"     "DMINORITY" "DEDUC"
[51] "DFEDTEN"  "DRETIRE"  "DDIS"     "DAGEGRP"  "DMIL"
[56] "DLEAVING"
```

Next, we remove rows with NA values. That deleted all rows with NA, empty string, space or “X”. Dataset is clean now.

```
> survey1 <- na.omit(survey)
```

Because the dataset is big, we randomly select 20k rows for analysis in this project.

```
> survey2 <- survey1[sample(nrow(survey1), 20000), ]
```

As introduced in last section, Satisfaction and Engagement are two most important parameters we are caring about. So, we need to add these two variables in the dataset. We use mean of the related question answers as the result, then minus 3 to make positive value indicate agree and negative value indicate disagree.

```
# Leaders Lead
> leaders <- (survey2$Q53 + survey2$Q54 + survey2$Q56 + survey2$Q60 + survey2$Q61)/5 -3
# Supervisors
> supervisors <- (survey2$Q47 + survey2$Q48 + survey2$Q49 + survey2$Q51 + survey2$Q52)/5
- 3
# Intrinsic Work Experience
> experience <- (survey2$Q3 + survey2$Q4 + survey2$Q6 + survey2$Q11 + survey2$Q12)/5 - 3
# Employee engagement index
> survey2$engagement <- (leaders + supervisors + experience)/3
# Global Satisfaction Index
> survey2$satisfaction <- (survey2$Q40 + survey2$Q69 + survey2$Q70 + survey2$Q71)/4 - 3
```

Then we need to transform the values of Satisfaction and Engagement to 1 and 0, in which 1 indicate agree and 0 indicate disagree or no idea. We use a user defined function to accomplish this task.

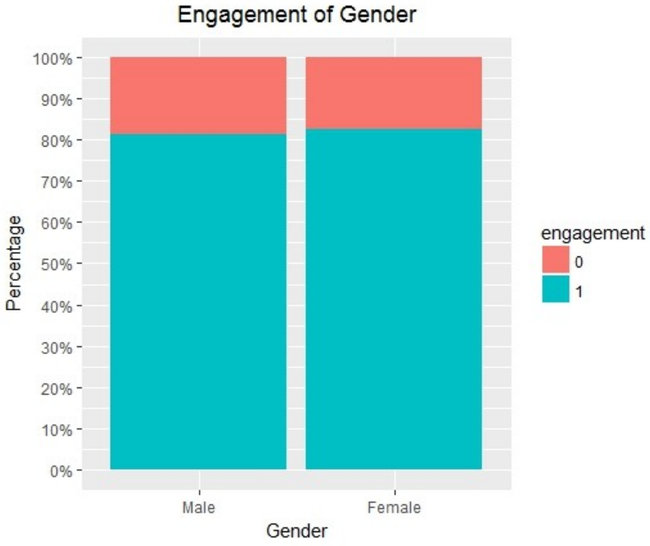
```
> positive <- function(v) {
+   u <- as.integer(sign(v))
+   for(i in 1:length(v))
+   {
+     if (v[i] <= 0) {
+       u[i] <- 0 }
+   }
+   return(u)
+ }
> survey3 <- survey2
> survey3$engagement <- as.factor(positive(survey2$engagement))
> survey3$satisfaction <- as.factor(positive(survey2$satisfaction))
```

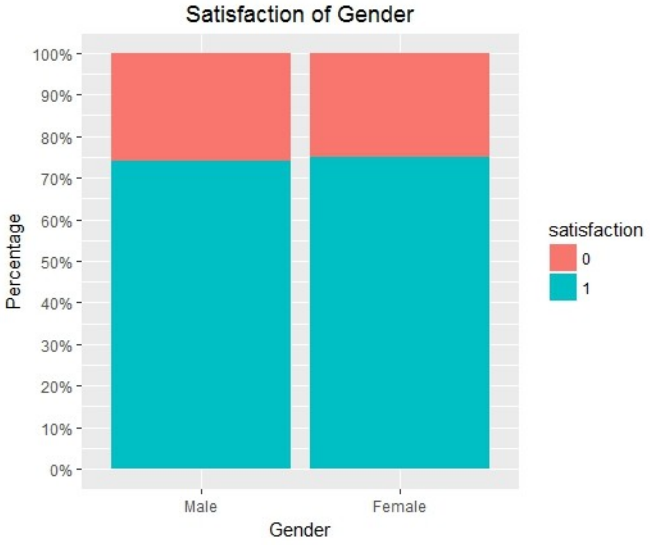
Now the dataset is ready for analysis, we can proceed to EDA stage.

V. Exploratory Data Analysis

In this stage, we'll do some comparison and visualization to find out the relationship among different groups and their engagement & satisfaction.

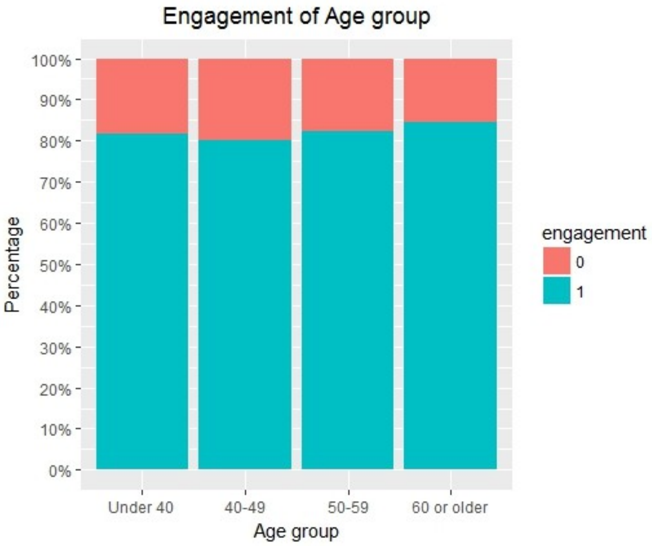
1. Gender

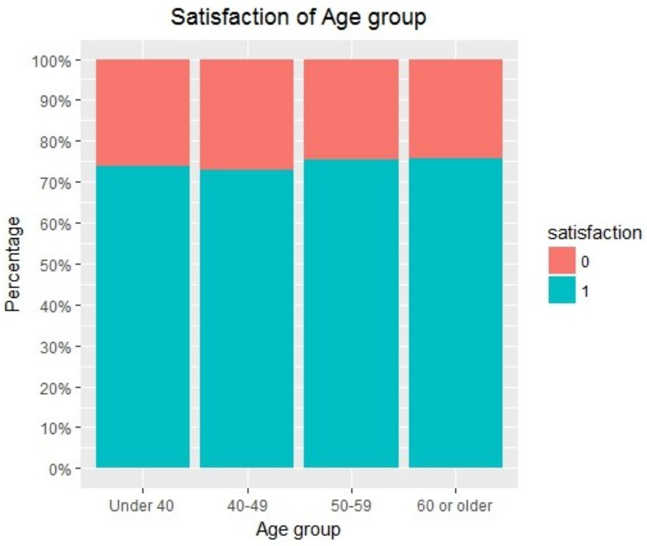




From the plots, we can see that there's no distinguish difference between males and females. Females are a little more engaged and satisfied than males.

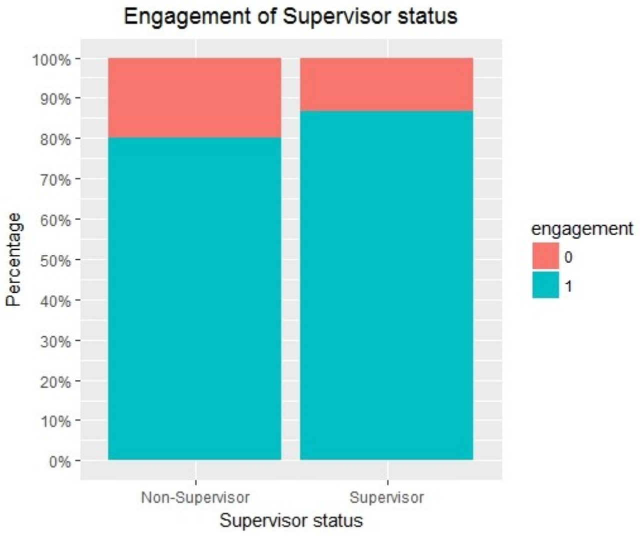
2. Age group

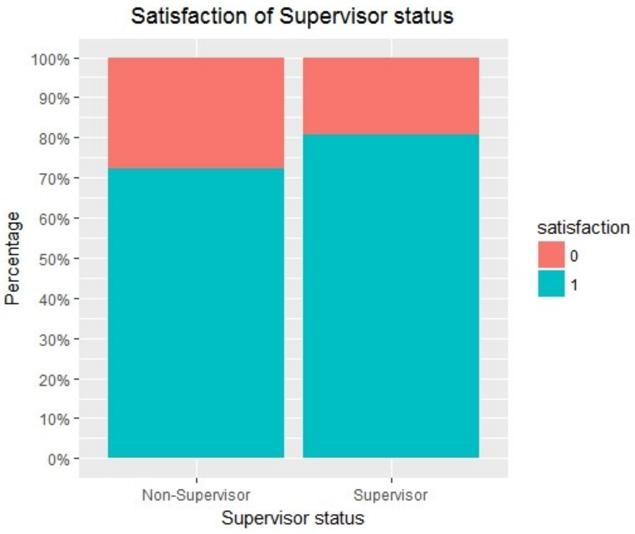




From the plots, we can see that there are some differences among age groups. Employees age 40-49 are least engaged and satisfied while employees over 60 are most engaged and satisfied with their organizations.

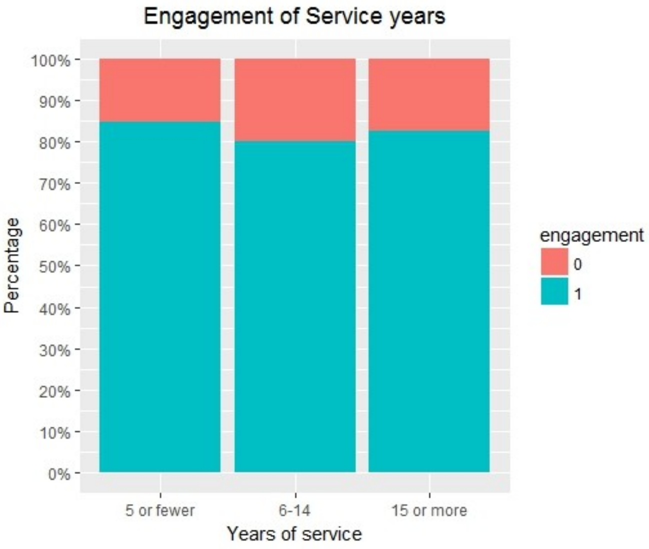
3. Supervisor status

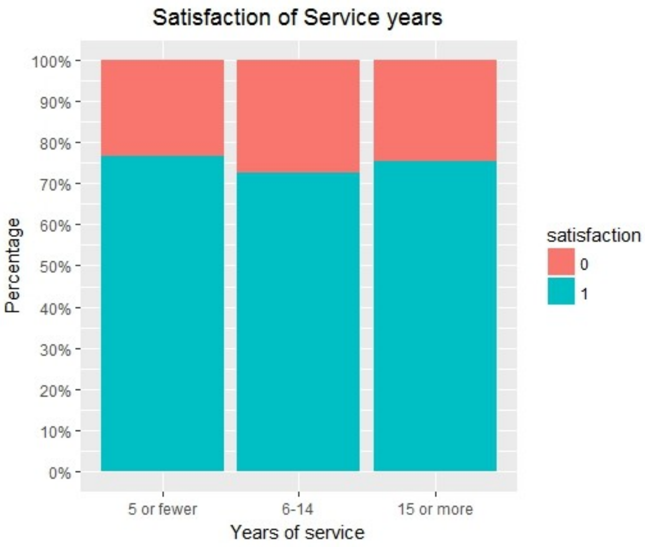




From the plots, we can see that there's distinguish difference between supervisors and non-supervisors. Supervisors are much more engaged and satisfied than non-supervisors. The differences are 6~8%.

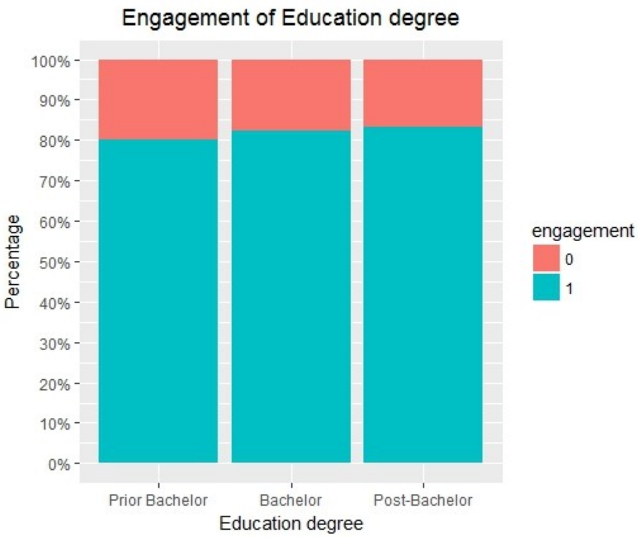
4. Years of service

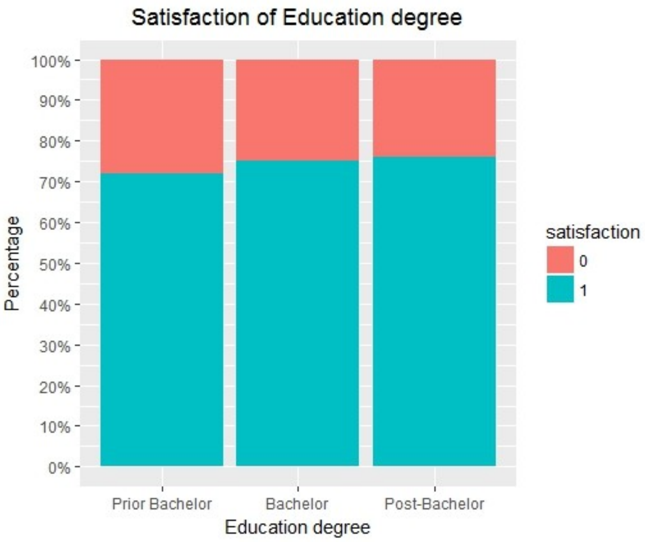

Engagement of Service years


Satisfaction of Service years

From the plots, we can see that there is some difference among service year groups. Employees with 6-14 years of service are much less engaged and satisfied than employees who worked fewer or more years.

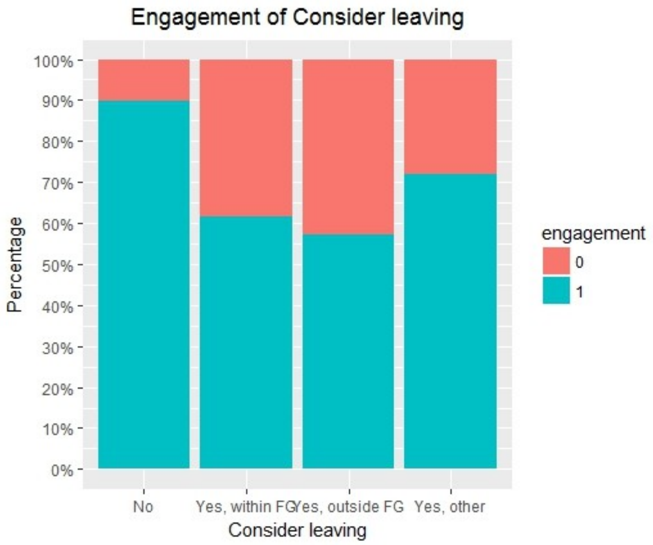
5. Education degree

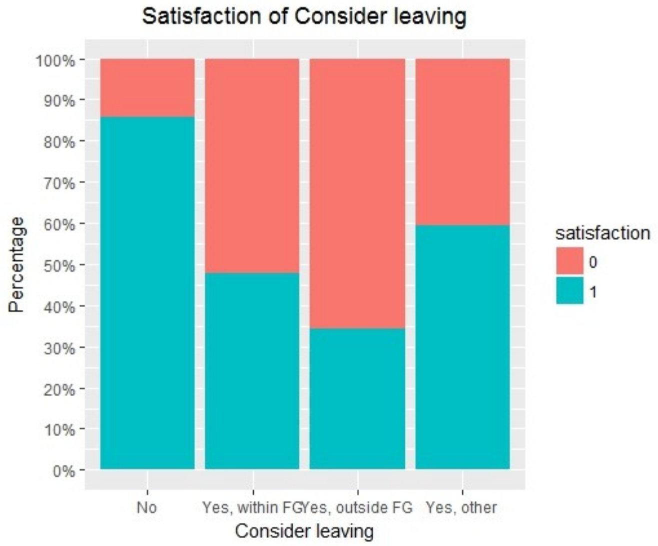

Engagement of Education degree


Satisfaction of Education degree

From the plots, we can see that there is a little difference among education degree groups. The higher the employee’ degree is, the more engaged and satisfied the employee feel. Through the difference is not distinguished.

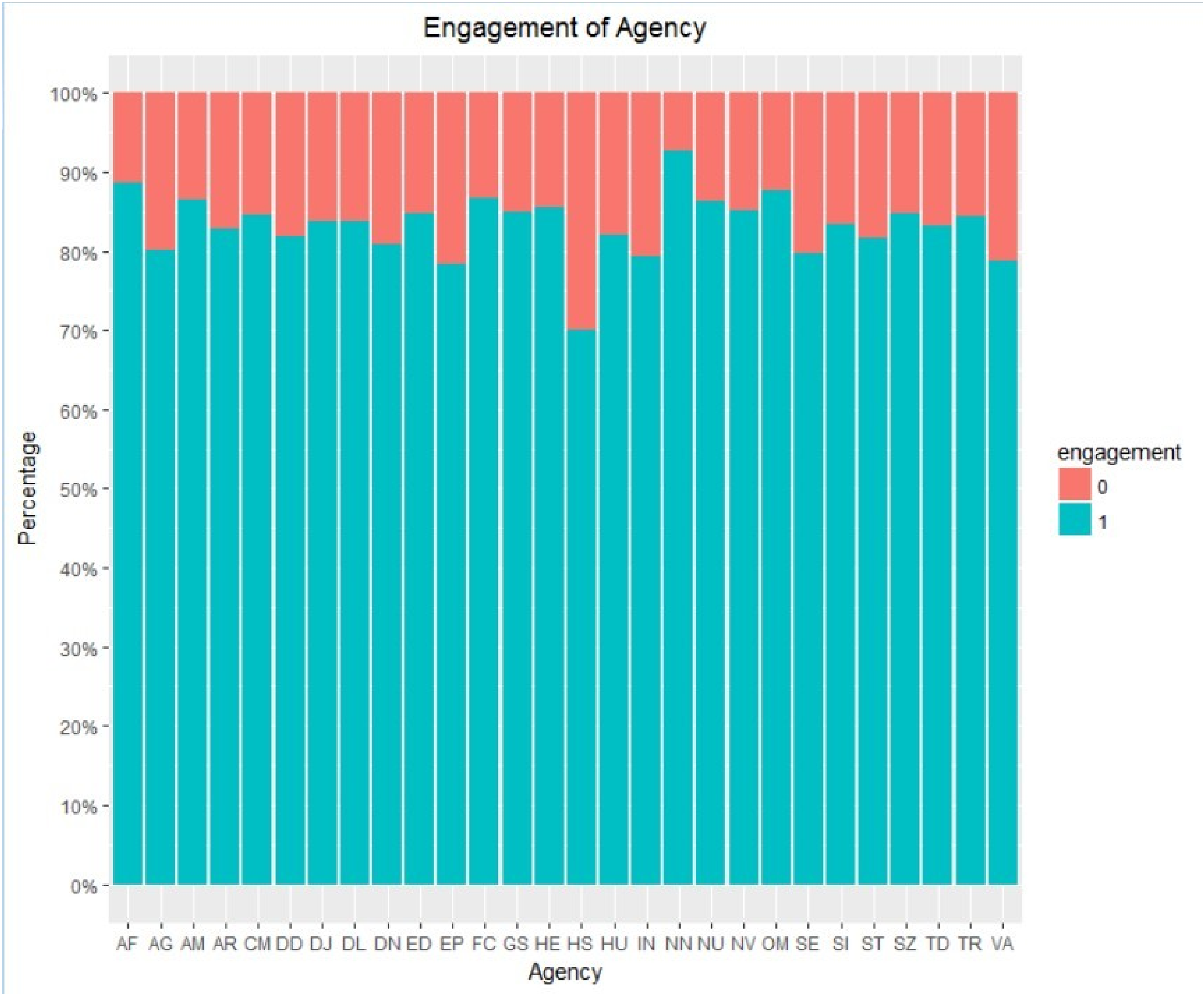
6. Consider leaving



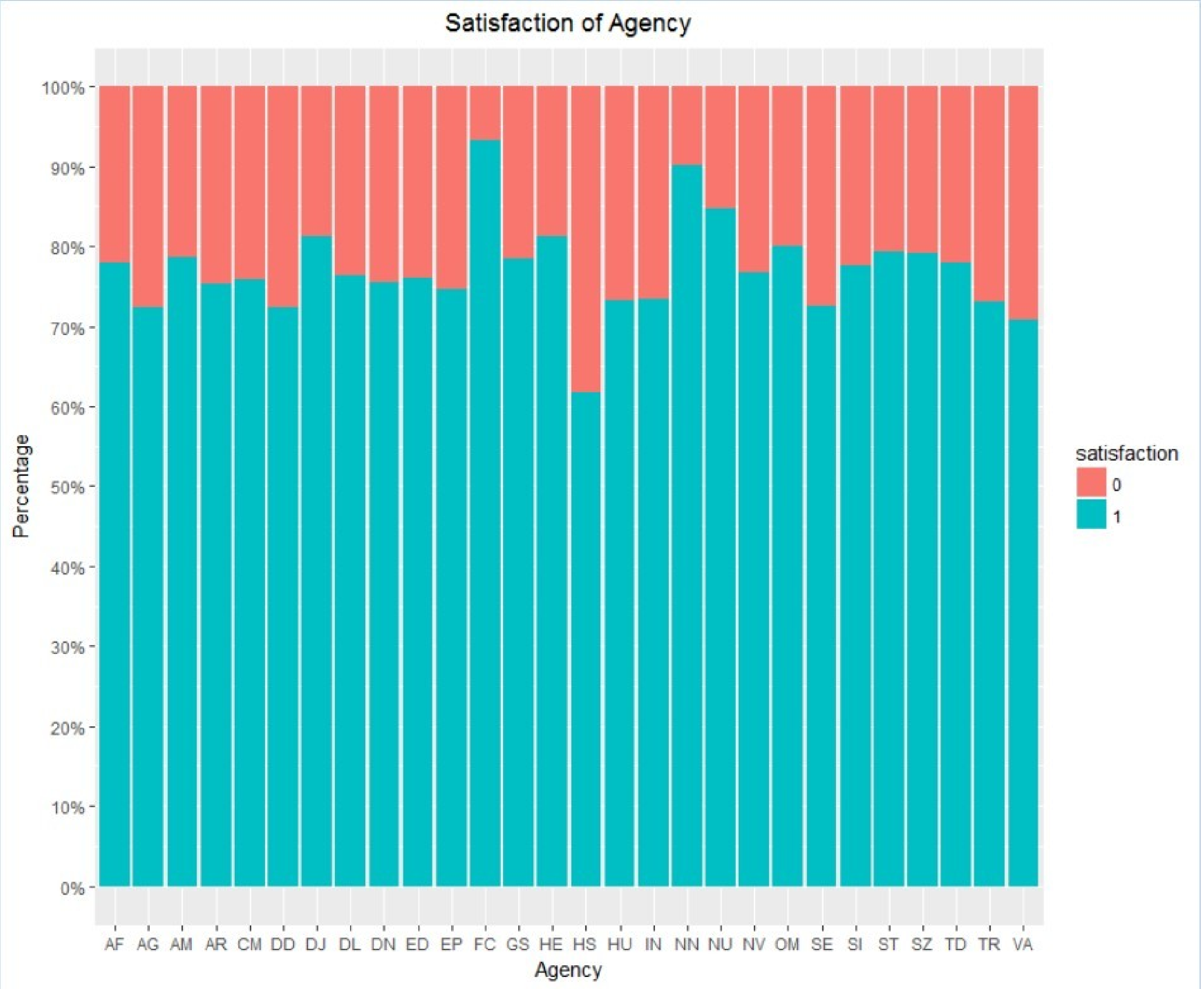


From the plots, we can see that there is big difference between employees who do and do not consider leaving. Employees don’t consider leaving are much more engaged and satisfied than employees who consider leaving, especially that consider to take another job outside the Federal Government. The difference is as big as 35~50%.

7. Agency



Engagement of different agencies are different. From the plot we know that HS(Department of Homeland Security) is the least engaged agency, and its engagement is about 70%. NN(National Aeronautics and Space Administration) is the most engaged agency, its engagement is about 95%. With this plot, leaders and managers of the agencies can get their agencies' engagement and consider if they need to work on improvement.



Satisfaction of different agencies are different. From the plot, we know that HS (Department of Homeland Security) is the least satisfied agency, and its satisfaction is about 62%. FC (Federal Communications Commission) is the most satisfied agency, its satisfaction is about 95% and NN (National Aeronautics and Space Administration) is the second satisfied agency, its satisfaction is about 90%. With this plot, leaders and managers of the agencies can get their agencies' satisfaction and consider if they need to work to make employees more satisfied.

VI. Data Modeling

In this step, we will find a model to predict the engagement of every employee by their survey answers and their demographics status.

1. Split dataset

First, we split the dataset into two data sets, 80%(16000 rows) is training data and 20%(4000 rows) is testing data. We’ll use training data to get the model and use testing data to measure effect of the model.

```
> set.seed(188)
# 80% data for training and 20% data for testing
> split = sample.split(survey3$engagement, SplitRatio = 0.8)
> surveyTrain = subset(survey3, split == TRUE)
> surveyTest = subset(survey3, split == FALSE)
> nrow(surveyTrain)
[1] 16000
> nrow(surveyTest)
[1] 4000
```

2. Build logistic regression model

Then we build logistic regression model to predict engagement. In the first model, we use all the variables.

```
> SurveyLog <- glm(engagement~., data=surveyTrain, family="binomial")
> summary(SurveyLog)
Call:
glm(formula = engagement ~ ., family = "binomial", data = surveyTrain)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.399   0.000   0.000   0.000   4.378

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.886e+03  9.142e+03  -0.206 0.836563
agencyAG              -9.093e+00  3.652e+00  -2.490 0.012790 *
agencyAM              -6.544e+01  3.301e+04  -0.002 0.998418
agencyAR              -1.945e+01  1.574e+02  -0.124 0.901641
agencyCM              -2.331e+00  4.131e+00  -0.564 0.572495
agencyDD              -1.803e+00  3.071e+00  -0.587 0.557049
agencyDJ             -1.877e+01  6.147e+02  -0.031 0.975636
agencyDL              -6.568e+00  3.208e+00  -2.047 0.040616 *
agencyDN              -2.760e+00  3.696e+00  -0.747 0.455190
agencyED              -2.268e+01  1.575e+04  -0.001 0.998851
agencyEP              -1.742e+01  9.720e+02  -0.018 0.985704
agencyFC              -5.636e+02  1.477e+05  -0.004 0.996956
agencyGS              -2.070e+01  1.087e+03  -0.019 0.984812
```

```
agencyHE              -2.392e+00  2.812e+00  -0.851 0.394837
agencyHS              -5.673e+00  3.093e+00  -1.834 0.066688 .
agencyHU              -2.197e+01  1.486e+03  -0.015 0.988205
agencyIN              -3.630e+00  2.852e+00  -1.273 0.203104
agencyNN              -7.299e+00  4.409e+00  -1.655 0.097855 .
agencyNU               1.989e+01  2.142e+04   0.001 0.999259
agencyNV              -6.199e+00  3.446e+00  -1.799 0.072048 .
agencyOM              -6.136e+01  2.659e+04  -0.002 0.998159
agencySE              -2.126e+01  2.413e+03  -0.009 0.992970
agencySI              -6.946e-01  2.758e+00  -0.252 0.801179
agencyST              -1.925e+01  1.021e+03  -0.019 0.984962
agencySZ               5.858e-01  2.771e+00   0.211 0.832587
agencyTD              -4.218e+00  3.073e+00  -1.373 0.169905
agencyTR              -5.256e+00  3.455e+00  -1.521 0.128173
agencyVA              -2.871e+00  2.670e+00  -1.075 0.282227
Q1                    -1.231e-01  5.012e-01  -0.246 0.806005
Q2                    -9.551e-01  4.681e-01  -2.041 0.041283 *
Q3                     4.265e+01  2.031e+02   0.210 0.833703
Q4                     4.271e+01  2.032e+02   0.210 0.833474
Q6                     4.285e+01  2.032e+02   0.211 0.832955
Q9                     3.911e-01  5.014e-01   0.780 0.435390
Q10                   -1.422e+00  5.377e-01  -2.644 0.008189 **
Q11                    4.207e+01  2.032e+02   0.207 0.835928
Q12                    4.315e+01  2.032e+02   0.212 0.831779
Q15                   -2.440e-01  5.610e-01  -0.435 0.663605
Q16                    5.425e-02  5.048e-01   0.107 0.914411
Q17                    3.743e-01  4.260e-01   0.879 0.379596
Q18                   -1.861e-02  4.786e-01  -0.039 0.968977
Q19                    8.261e-01  4.447e-01   1.858 0.063229 .
Q22                   -5.433e-01  6.085e-01  -0.893 0.371995
Q23                   -6.225e-01  5.252e-01  -1.185 0.235903
Q24                    1.064e+00  8.622e-01   1.234 0.217132
Q25                   -7.860e-01  5.847e-01  -1.344 0.178878
Q30                    5.581e-02  5.821e-01   0.096 0.923619
Q32                    5.985e-01  6.829e-01   0.876 0.380837
Q34                    1.222e+00  5.405e-01   2.260 0.023827 *
Q37                    1.460e+00  5.536e-01   2.637 0.008363 **
Q38                   -1.440e+00  5.571e-01  -2.585 0.009741 **
Q40                   -2.029e+00  6.857e-01  -2.959 0.003089 **
Q42                    1.146e+00  6.152e-01   1.863 0.062506 .
Q44                    6.062e-01  6.451e-01   0.940 0.347351
```

```
Q45                      -3.320e-01  6.372e-01  -0.521 0.602378
Q46                      -5.364e-01  6.040e-01  -0.888 0.374502
Q47                       4.211e+01  2.032e+02   0.207 0.835780
Q48                       4.240e+01  2.031e+02   0.209 0.834655
Q49                       4.044e+01  2.031e+02   0.199 0.842196
Q50                       5.437e-02  5.317e-01   0.102 0.918559
Q51                       4.055e+01  2.031e+02   0.200 0.841765
Q52                       4.308e+01  2.031e+02   0.212 0.832061
Q53                       4.149e+01  2.031e+02   0.204 0.838155
Q54                       4.060e+01  2.031e+02   0.200 0.841578
Q55                       8.606e-01  5.577e-01   1.543 0.122804
Q56                       4.147e+01  2.031e+02   0.204 0.838258
Q58                       2.167e+00  8.753e-01   2.475 0.013309 *
Q59                      -2.653e+00  7.897e-01  -3.359 0.000782 ***
Q60                       4.163e+01  2.031e+02   0.205 0.837647
Q61                       4.243e+01  2.032e+02   0.209 0.834547
Q69                      -1.086e+00  6.067e-01  -1.790 0.073377 .
Q70                       8.416e-01  4.163e-01   2.021 0.043232 *
Q71                      -1.286e-02  5.902e-01  -0.022 0.982622
DSUPERSupervisor          1.415e-01  1.387e+00   0.102 0.918726
DSEXFemale               -2.274e-01  1.089e+00  -0.209 0.834564
DMINORITY                 2.167e+00  1.191e+00   1.819 0.068874 .
DEDUCBachelor             1.313e+00  9.894e-01   1.327 0.184453
DEDUCPost-Bachelor       -4.786e-01  1.295e+00  -0.370 0.711718
DFEDTEN6-14              -2.641e+00  1.414e+00  -1.867 0.061896 .
DFEDTEN15 or more        -1.070e+00  1.386e+00  -0.772 0.440115
DRETIREB                 -1.304e+00  1.496e+00  -0.872 0.383457
DDISB                    -2.240e+00  1.020e+00  -2.196 0.028070 *
DAGEGRP40-49              3.135e-01  1.827e+00   0.172 0.863765
DAGEGRP50-59             -1.932e+00  2.010e+00  -0.961 0.336405
DAGEGRP60 or older       -4.529e+00  2.899e+00  -1.562 0.118205
DMILB                     8.763e+00  3.052e+00   2.871 0.004092 **
DMILC                     8.841e-01  1.615e+00   0.548 0.584011
DMILD                    -1.131e+00  1.289e+00  -0.877 0.380401
DLEAVINGYes, within FG   -2.340e+00  1.387e+00  -1.687 0.091597 .
DLEAVINGYes, outside FG  -7.740e-01  1.725e+00  -0.449 0.653626
DLEAVINGYes, other       -1.416e+00  1.494e+00  -0.948 0.343210
satisfaction1             2.659e+00  1.410e+00   1.886 0.059362 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15148.09  on 15999  degrees of freedom
Residual deviance:   111.92  on 15908  degrees of freedom
AIC: 295.92

Number of Fisher Scoring iterations: 25
```

In the second model, we only use the significant variables for the first model.

```
> SurveyLog1 <- glm(engagement ~ agency + Q2 + Q10 + Q19 + Q34 + Q37 + Q38 + Q40 + Q42
+ Q58 + Q59 + Q69 + Q70 + DMINORITY + DFEDTEN + DDIS + DMIL + DLEAVING + satisfaction,
data=surveyTrain, family="binomial")
> summary(SurveyLog1)
Call:
glm(formula = engagement ~ agency + Q2 + Q10 + Q19 + Q34 + Q37 +
    Q38 + Q40 + Q42 + Q58 + Q59 + Q69 + Q70 + DMINORITY + DFEDTEN +
    DDIS + DMIL + DLEAVING + satisfaction, family = "binomial",
    data = surveyTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3281   0.0150   0.0718   0.2001   3.8898

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -11.894621   0.440210 -27.020  < 2e-16 ***
agencyAG              -1.016653   0.240105  -4.234 2.29e-05 ***
agencyAM              -0.406180   0.627147  -0.648 0.517203
agencyAR              -0.664937   0.241125  -2.758 0.005822 **
agencyCM              -0.387211   0.323064  -1.199 0.230700
agencyDD              -0.611157   0.264658  -2.309 0.020931 *
agencyDJ              -0.744597   0.258228  -2.883 0.003933 **
agencyDL              -0.287493   0.306810  -0.937 0.348739
agencyDN              -0.770757   0.317788  -2.425 0.015292 *
agencyED              -0.001044   0.503580  -0.002 0.998346
agencyEP              -0.998532   0.392021  -2.547 0.010861 *
agencyFC              -1.720344   2.003250  -0.859 0.390464
agencyGS              -0.681460   0.329724  -2.067 0.038757 *
agencyHE              -0.709256   0.237608  -2.985 0.002836 **
```

```
agencyHS              -0.597397   0.210177  -2.842 0.004478 **
agencyHU              -0.574945   0.403989  -1.423 0.154687
agencyIN              -0.882008   0.237367  -3.716 0.000203 ***
agencyNN              -0.678240   0.426095  -1.592 0.111439
agencyNU              -0.603965   0.668755  -0.903 0.366463
agencyNV              -0.307475   0.259674  -1.184 0.236381
agencyOM               0.157892   0.651568   0.242 0.808528
agencySE               0.383097   0.767427   0.499 0.617641
agencySI              -0.936489   0.284448  -3.292 0.000994 ***
agencyST              -1.516199   0.394265  -3.846 0.000120 ***
agencySZ              -0.221828   0.327986  -0.676 0.498829
agencyTD              -0.859269   0.267210  -3.216 0.001301 **
agencyTR              -0.694192   0.224519  -3.092 0.001989 **
agencyVA              -0.568660   0.227708  -2.497 0.012513 *
Q2                     0.520396   0.040655  12.800  < 2e-16 ***
Q10                   -0.049744   0.034259  -1.452 0.146498
Q19                    0.495224   0.032697  15.146  < 2e-16 ***
Q34                    0.098035   0.039229   2.499 0.012451 *
Q37                    0.385673   0.039647   9.728  < 2e-16 ***
Q38                    0.186201   0.042778   4.353 1.34e-05 ***
Q40                    0.333188   0.052018   6.405 1.50e-10 ***
Q42                    0.861113   0.039791  21.641  < 2e-16 ***
Q58                    0.649053   0.056211  11.547  < 2e-16 ***
Q59                    0.396620   0.055047   7.205 5.80e-13 ***
Q69                    0.658311   0.052055  12.646  < 2e-16 ***
Q70                   -0.090953   0.038306  -2.374 0.017580 *
DMINORITY             -0.161555   0.081824  -1.974 0.048333 *
DFEDTEN6-14           -0.059692   0.109188  -0.547 0.584592
DFEDTEN15 or more     -0.088036   0.111858  -0.787 0.431261
DDISB                 -0.282093   0.114475  -2.464 0.013731 *
DMILB                 -0.264829   0.254786  -1.039 0.298610
DMILC                 -0.495770   0.136403  -3.635 0.000278 ***
DMILD                 -0.426838   0.100247  -4.258 2.06e-05 ***
DLEAVINGYes, within FG   -0.146750   0.092558  -1.585 0.112855
DLEAVINGYes, outside FG  -0.044211   0.170069  -0.260 0.794895
DLEAVINGYes, other        0.016419   0.113091   0.145 0.884567
satisfaction1          0.525925   0.115686   4.546 5.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 15148.1  on 15999  degrees of freedom
Residual deviance:  5000.2  on 15949  degrees of freedom
AIC: 5102.2

Number of Fisher Scoring iterations: 7
```

From the summary of the model, we can see that variables are more significant than that in the first model. But AIC is much bigger. Then in the third model, we remove agency and satisfaction from variable list.

```
> SurveyLog2 <- glm(engagement ~ Q2 + Q10 + Q19 + Q34 + Q37 + Q38 + Q40 + Q42 + Q58 +
Q59 + Q69 + Q70 + DMINORITY + DFEDTEN + DDIS + DMIL + DLEAVING, data=surveyTrain,
family="binomial")
> summary(SurveyLog2)
Call:
glm(formula = engagement ~ Q2 + Q10 + Q19 + Q34 + Q37 + Q38 +
    Q40 + Q42 + Q58 + Q59 + Q69 + Q70 + DMINORITY + DFEDTEN +
    DDIS + DMIL + DLEAVING, family = "binomial", data = surveyTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2709   0.0140   0.0736   0.2072   4.1328

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)           -13.066617   0.365855 -35.715  < 2e-16 ***
Q2                      0.527915   0.040307  13.097  < 2e-16 ***
Q10                    -0.045428   0.033898  -1.340  0.18021
Q19                     0.492417   0.032264  15.262  < 2e-16 ***
Q34                     0.084066   0.038753   2.169  0.03006 *
Q37                     0.381839   0.039012   9.788  < 2e-16 ***
Q38                     0.194623   0.042440   4.586 4.52e-06 ***
Q40                     0.422032   0.046350   9.105  < 2e-16 ***
Q42                     0.852599   0.039355  21.665  < 2e-16 ***
Q58                     0.656762   0.055836  11.762  < 2e-16 ***
Q59                     0.398478   0.054798   7.272 3.55e-13 ***
Q69                     0.743347   0.047379  15.689  < 2e-16 ***
Q70                    -0.009958   0.033470  -0.298  0.76606
DMINORITY              -0.175630   0.080064  -2.194  0.02826 *
DFEDTEN6-14            -0.059783   0.107305  -0.557  0.57744
```

```
DFEDTEN15 or more       -0.128463   0.108824  -1.180  0.23781
DDISB                  -0.289294   0.112760  -2.566  0.01030 *
DMILB                  -0.072937   0.248571  -0.293  0.76920
DMILC                  -0.329935   0.125514  -2.629  0.00857 **
DMILD                  -0.391668   0.097283  -4.026 5.67e-05 ***
DLEAVINGYes, within FG   -0.140979   0.090708  -1.554  0.12014
DLEAVINGYes, outside FG  -0.071682   0.168004  -0.427  0.66962
DLEAVINGYes, other       0.003858   0.112169   0.034  0.97256
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15148.1  on 15999  degrees of freedom
Residual deviance:  5067.1  on 15977  degrees of freedom
AIC: 5113.1

Number of Fisher Scoring iterations: 7
```

This time AIC is getting a little bigger than that of model 2. So, compare these 3 model, the first model is the best one. We’ll use it in following analysis.

3. Predict training dataset

```
> predictTrain = predict(SurveyLog, type="response")
> summary(predictTrain)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.0000  1.0000  1.0000  0.8187  1.0000  1.0000 
> tapply(predictTrain, surveyTrain$engagement, mean)
          0           1 
0.005569868 0.998766456 
```

The tapply result means that we predict 99.88% ‘1’ as ‘1’ and 0.56% ‘0’ as ‘1’. This is a very precise model.

```
> predictTrain1 = predict(SurveyLog1, type="response")
> summary(predictTrain1)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0000647 0.8336000 0.9896000 0.8187000 0.9987000 1.0000000 
> tapply(predictTrain1, surveyTrain$engagement, mean)
        0         1 
0.2601498 0.9423853 
```

```
> predictTrain2 = predict(SurveyLog2, type="response")
> summary(predictTrain2)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0000647 0.8331000 0.9890000 0.8187000 0.9986000 1.0000000 
> tapply(predictTrain2, surveyTrain$engagement, mean)
        0         1 
0.2630620 0.9417404 
```

The other two models predict about 94% ‘1’ as ‘1’ and 26% ‘0’ as ‘1’, they are much less precise than the first model. So, we’ll use the first model in following analysis.

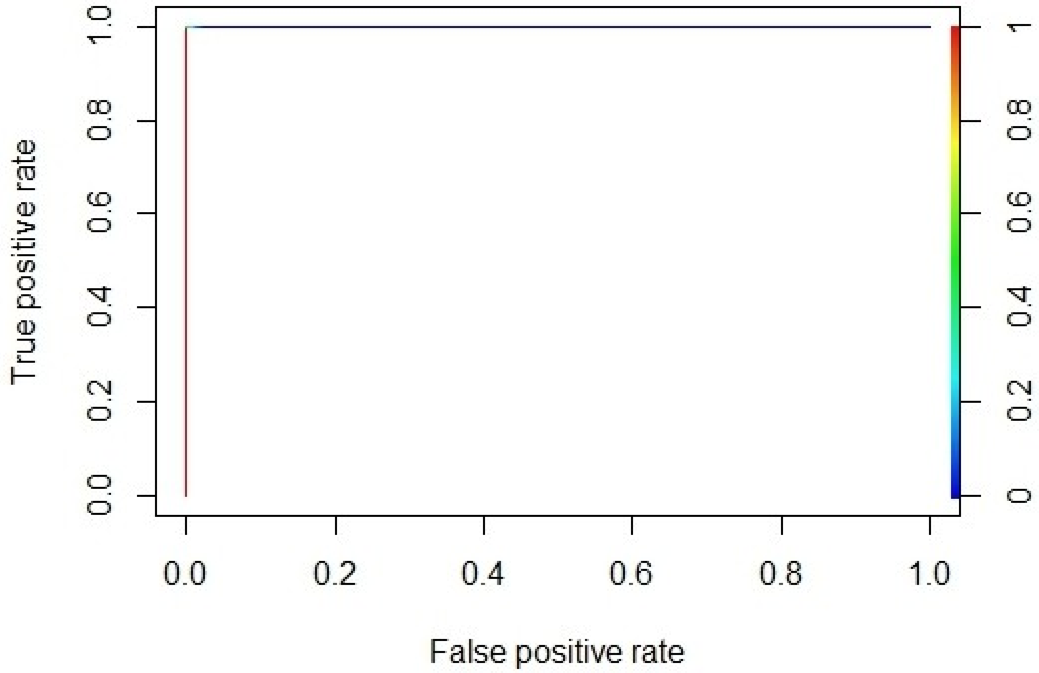
4. Confusion matrix (compare predicted vs actual)

```
> table(surveyTrain$engagement, predictTrain > 0.5)
   FALSE  TRUE
  0  2895     6
  1    11 13088
> 13088/13099                  # Sensitivity
[1] 0.9991602
> 2895/2901                    # Specificity
[1] 0.9979317
> table(surveyTrain$engagement, predictTrain > 0.7)
   FALSE  TRUE
  0  2901     0
  1    25 13074
> 13074/13099
[1] 0.9980915
> 2901/2901
[1] 1
> table(surveyTrain$engagement, predictTrain > 0.2)
   FALSE  TRUE
  0  2868    33
  1     4 13095
> 13095/13099
[1] 0.9996946
> 2868/2901
[1] 0.9886246
```

We tried 0.5, 0.7 and 0.2 as the threshold respectively, Sensitivities are all more than 99.8% and Specificities have a little difference from 98.86% to 1. Through comprehensive evaluation, 0.5 may be the best threshold.

5. ROC curve

```
> ROCRpred = prediction(predictTrain, surveyTrain$engagement)
> ROCRperf = performance(ROCRpred, "tpr", "fpr")
> plot(ROCRperf, colorize=TRUE)
```



6. Predict test dataset with the model

```
> predictTest = predict(SurveyLog, type="response", newdata=surveyTest)
> summary(predictTest)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  1.0000  1.0000  0.8189  1.0000  1.0000 
> tapply(predictTest, surveyTest$engagement, mean)
         0          1 
0.01188826 0.99754287 
```

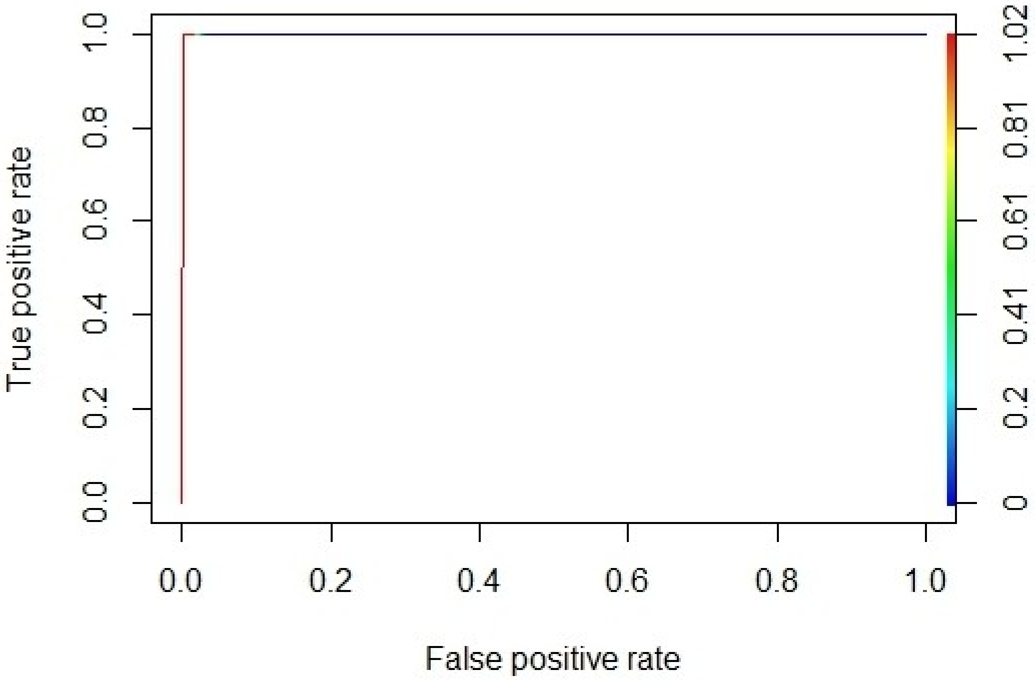
This result on testing dataset means that we predict 99.75% ‘1’ as ‘1’ and 1.19% ‘0’ as ‘1’. This is a very good result.

```
> table(surveyTest$engagement, predictTest > 0.5)
   
    FALSE TRUE
  0   717    8
  1     8 3267
> 3267/3275        # Sensitivity
[1] 0.9975573
> 717/725          # Specificity
```

```
[1] 0.9889655
> table(surveyTest$engagement, predictTest > 0.7)
   FALSE TRUE
  0   721    4
  1     8 3267
> 3267/3275        # Sensitivity
[1] 0.9975573
> 721/725          # Specificity
[1] 0.9944828
> table(surveyTest$engagement, predictTest > 0.3)
   FALSE TRUE
  0   714   11
  1     8 3267
> 3267/3275
[1] 0.9975573
> 714/725
[1] 0.9848276
```

We tried 0.5, 0.7 and 0.3 as the threshold respectively. Through comprehensive evaluation, 0.7 may be the best threshold.

```
> ROCRpredTest = prediction(predictTest, surveyTest$engagement)
> auc = as.numeric(performance(ROCRpredTest, "auc")@y.values)
> auc
[1] 0.9998425
> ROCRperfTest = performance(ROCRpredTest, "tpr", "fpr")
> plot(ROCRperfTest, colorize=TRUE)
```



We have applied the model in the test dataset and the predict result is very good. auc=0.9998425 means the model is very accurate.

VII. Results and Discussion

In this project, we did some research and analysis about the answers of the Federal Employee Viewpoint Survey (FEVS). Because engagement and satisfaction are the most important measures of employees' attitude toward their organization, we tried to find the relationship among employees' status and their engagement and satisfaction.

From the research and analysis, we can get the following results:

1. There's no distinguish different feeling between males and females
2. There are some differences among age groups. Employees age 40-49 are least engaged and satisfied while employees over 60 are most engaged and satisfied with their organizations. This may because employees age 40-49 are the most talented ones and in good shape, so they have more choices in job seeking.
3. There's distinguish difference between supervisors and non-supervisors. Supervisors are much more engaged and satisfied with their organizations than non-supervisors.

4. There is some difference among service year groups. Employees with 6-14 years of service are much less engaged and satisfied than employees who worked fewer or more years.
5. There is a little difference among education degree groups. The higher the employee' degree is, the more engaged and satisfied the employee feel. But the difference is not distinguished.
6. There is big difference between employees who do and do not consider leaving. Employees don't consider leaving are much more engaged and satisfied than employees who consider leaving, especially that consider to take another job outside the Federal Government. The difference is as big as 35~50%.
7. Employees in different agencies have different engagement and satisfaction.

We made a model through logistic regression to analyze and predict the engagement. From the summary of the model, we can get some information on which aspects affect engagement. Questions Q2, Q19, Q34, Q37, Q38, Q40, Q42, Q58, Q59, Q69, Q70 have significant correlation with engagement, those indicate the following aspects respectively:

1. Job Resources: allow sufficient materials, knowledge, personnel, skills, information and work distribution to complete the job
2. Performance Rating: ensure employees are held accountable and performance is evaluated and rated
3. Merit System Principles: support fairness and protect employees from arbitrary actions, favoritism, political coercion, and reprisal
4. Collaborative Management: promote and support collaborative communication and teamwork in accomplishing goals and objectives
5. Work/Life Balance: support employee needs to balance work and life responsibilities
6. Global Satisfaction: employees are satisfied with their job and pay, they would recommend their organization as a good place to work
7. Open: Does management support diversity in all ways?

We may call these aspects as the engagement drivers. Leaders and managers can work on these aspects to improve the engagement of their employees.

Finally, we built a very good model with 80% of the sample data to predict every employee's engagement from their survey answers. This model turned out to be very accurate, we can use it on the large data outside sample dataset.

VIII. Future Work

We did research and analysis mainly about single employee's answers and their results in this project. In the future, more jobs can be done to get the engagement and satisfaction of each work unit.

1. Engagement of work units
2. Make a model to predict the engagement of a work unit
3. Do research on how to improve the engagement of a work unit
4. Compare the results of different years to find out if there are some improvement on different aspects.

Reference

1. 2015_FEVS_Gwide_Final_Report
   https://www.fedview.opm.gov/2015FILES/2015_FEVS_Gwide_Final_Report.pdf
2. FEVS_Engagement_INFOGRAPHIC-2015
   https://www.fedview.opm.gov/2015FILES/FEVS_Engagement_INFOGRAPHIC.pdf
3. Engagement_Drivers_Background_and_Summary
   https://www.fedview.opm.gov/2015FILES/Engagement_Drivers_Background_and_Summary.pdf
4. 2015-Trends-in-Global-Employee-Engagement-Report
   http://www.aon.com/attachments/human-capital-consulting/2015-Trends-in-Global-Employee-Engagement-Report.pdf