Experiments for Business Analytics

Yashovardhan Bhomia

200680497

MSc Business Analytics

Introduction

In the new era where transport cost is increasing, traffic has increased which has led to congestion on roads and people need to commute long distances to go to offices, demand of Working from Home has increased. After World has faced Covid pandemic, the word for Working from Home has increased drastically. In United States, the proportion of employees who do work from home has been tripled in last 3 decades. (Mateyka, 2010)

In International market, working from home is the new normal way of living. Studies have been conducted to see whether how many managers are allowed to do work from home, interestingly, United States, Germany and United Kingdom data shows that almost 50% of managers are working from home and at the same time, developing countries also see a rise in this.

There can be numerous factors which affects employee's satisfaction level in a company. So, let's analyze such factors by conducting experiment on satisfaction data from Work from Home experiment conducted at Ctrip, China's largest travel agency, having 16000 employees and is listed on NASDAQ. Management was interested to allow work from home to employees to reduce office rental costs, reduce attrition rates and increase satisfaction among employees. (Liang, 2014)

So, idea is to run an experiment to see whether there is an impact on Satisfaction factor by taking data from a firm when employees work from home.

This report will measure impact on satisfaction with variables such as age, tenure, high education, and volunteer using data of Shanghai Ctrip call Centre. Through linear and logistic regression, experiment shows that age and tenure have no effect on satisfaction, but Higher education and volunteer has an impact on satisfaction.

This experiment will cover the multiple sections, second section will cover the Main question, later discussion around Hypothesis will be covered. Hypothesis is based on age, tenure, Higher education, and Volunteer. In fourth section, tabular and graphical description is used to explore the data. In last section, it will contain and explain the results of the analysis done based on different models.

Question

Ctrip conducted experiment for 9 months on working from home, and collected different type of observations like performance, attrition, exhaustion, satisfaction etc.

Main question of this report is to check impact on satisfaction scores due to various aspects and variables. But before developing the hypothesis, lets deep dive in defining satisfaction and discuss the same. Satisfaction means "fulfillment of one's wishes, expectations and needs". In today's competitive business market, it's very essential that employees feel satisfied in the work they do or how they approach life. So, it's a very urgent problem which needs to be studied, as job and life satisfaction plays a key role in performance of an employee.

Question – Do Satisfaction level is impacted by other external factors?

In previous literature (Bloom), there were some key findings, employees who worked from home reported substantially higher work satisfaction, had more positive outcomes in additional survey. Attrition also decreased by 50% for employees doing WFH compared to control group. Overall effect of this experiment was that Ctrip saved $2000 a year per employee and productivity increased by almost 30%.

Satisfaction is also impacted by how big office infrastructure is and how office is run functionally by management. Lowest satisfaction is generally reported in Congested office structures and properly planned offices with big workplace area have higher satisfaction among employees. (Bodin, 2008)

Hypotheses and Methods

Ctrip senior management was interested to know about self-reported satisfaction levels of employee and how it was affected by the experiment. So basically, they ran two sets of surveys, a satisfaction survey and a work attitude survey.
It was done conducting tests developed by psychologists in 1980's. The Satisfaction survey was conducted total 5 times for each employee, 1st survey was done before randomization was done and other 4 were done during the experiment. Initially there was not much difference between survey results of treatment and control group, but as soon as experiment started, employees in treatment group started reporting higher satisfaction levels.

So, it is very important to see which factors contributed to employees from treatment group giving higher satisfaction levels. This leads to generation of 2 hypothesis which this report will be testing.

Hypothesis 1 - Satisfaction is higher for employees from treatment group having higher age and tenure in company.

$$\text{Satisfaction} = B0*\text{age} + B1*\text{tenure}$$

In general, if a person joins a company at a young age and continues with company for many years, it can be inferred that employee must be having high level of satisfaction level, as he has not quit the job even after working for many years with same company. It will be interesting to see that employee having high age and high tenure have more satisfaction level.

Hypothesis 2 - Satisfaction is lower for employees who are working from home and have high education and do volunteer work for society.

Satisfaction is also impacted by how well educated a employee is, employee having higher education degree has more ambitious lifestyle, and also if he indulges in volunteer work.

Satisfaction = B0*high_educ + B1*volunteer

After setting up above hypothesis, the variables are divided into 3 main categories: -

(a) Dependent Variable - To answer above mentioned hypothesis, “satisfaction” will the dependent or response variable, which is affected by independent variables.
(b) Independent Variables - age, tenure, high_educ and volunteer will be the independent variables.
(c) Fixed affect Variables - survey no and person id are fixed affect variables, so they need to be omitted while results of regression models.

This report will use Ordinary Least Squares (OLS) which is a linear regression model to do preliminary regression. Since experiment divides satisfaction levels into two categories (low and high), results of OLS might not be accurate, so logistic regression will also be used.

Then the coefficient estimates will be monitored to see whether there is a positive, negative or no impact of independent variables over dependent variable. Marginal effects will also be used to derive meaningful quantities from regression estimates.

Data Description

Before creating different regression models, let’s explore Satisfaction Dataset. This dataset is collected from Working from Home experiment in Ctrip from 2010 to 2011.

171 employees participated in Satisfaction based survey. In Table 1, the brief information of the dataset is given. There is total 855 observations ,17 variables and there are no missing values. In addition, the dependent variable, independents, and controlled variables are recorded in it.

Table 2 displays that Satisfaction variable has range of values from 1 to 7 and its states that mostly Satisfaction level 5 has been selected by employees.

Table 1 : Dataset Information

```
===========================================================================
Statistic           N      Mean      St. Dev.    Min  Pctl(25) Pctl(75)  Max
---------------------------------------------------------------------------
surveyno           855   3.000       1.415       1       2        4       5
satisfaction       855   4.695       1.352       1       4        6       7
general            855   72.775      11.742     32     64.5      80      100
life               855   21.460      7.353       2      16       27      38
personid           855 29,094.220 12,042.630 3,906  18,112   39,512  45,442
expgroup_treatment 855   0.501       0.500       0       0        1       1
age                855   24.661      3.623      18      22       27      35
tenure             855   28.289      22.618      2       9       45      96
grosswage          855   3.076       0.810     1.388  2.500    3.637   6.221
children           855   0.175       0.381       0       0        0       1
bedroom            855   0.971       0.169       0       1        1       1
commute            855  110.939      62.413      2      60      180      300
men                855   0.468       0.499       0       0        1       1
married            855   0.263       0.441       0       0        1       1
volunteer          855   0.877       0.328       0       1        1       1
high_educ          855   0.357       0.479       0       0        1       1
T_pid              855   0.200       0.400       0       0        0       1
___________________________________________________________________________
```

Table 2 : Satisfaction by Person ID

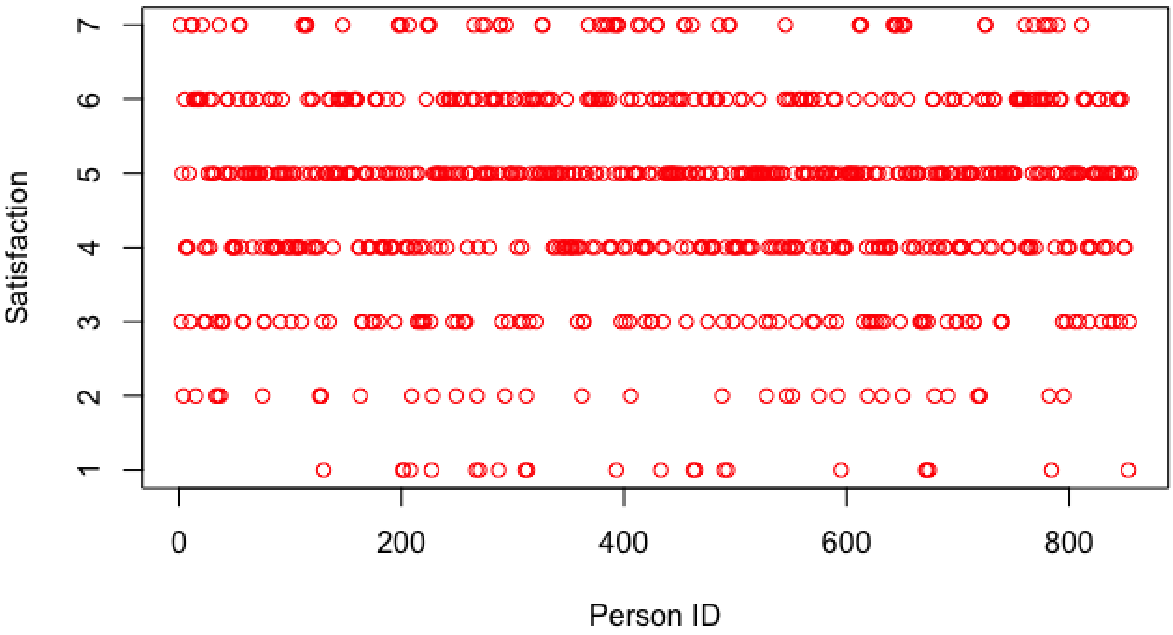


Table 3 shows the Age Distribution in Company, which is between 18 and 35. There are a lot of young employees in company as largest number of data is present between Age 20 to Age 25. Also, there are very less employees as age increases.

Table 4 states the tenure against number of employees. It can be interpreted that majority of the employees have less than 40 months of tenure with company and for Age group 45 to 60, tenure of the employees is very less

Table 3 : Age Distribution in Company

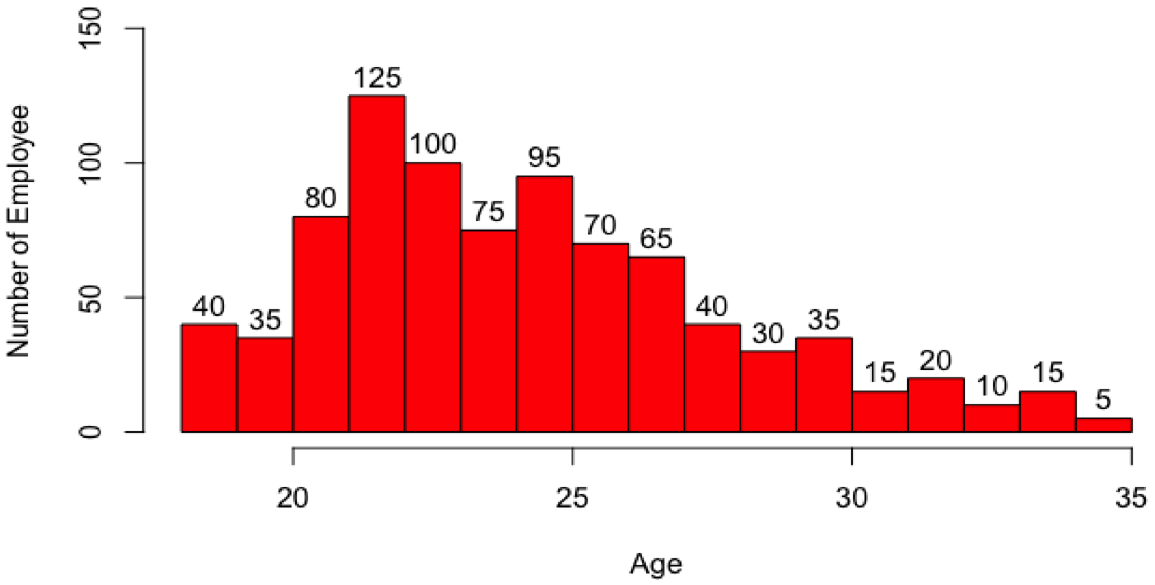


Table 4 : Tenure Distribution in Company

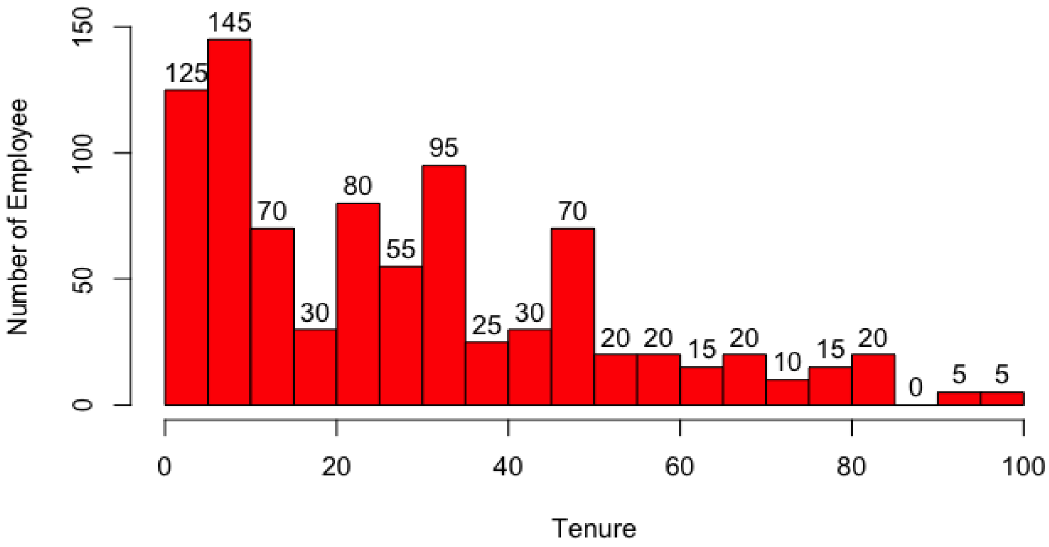
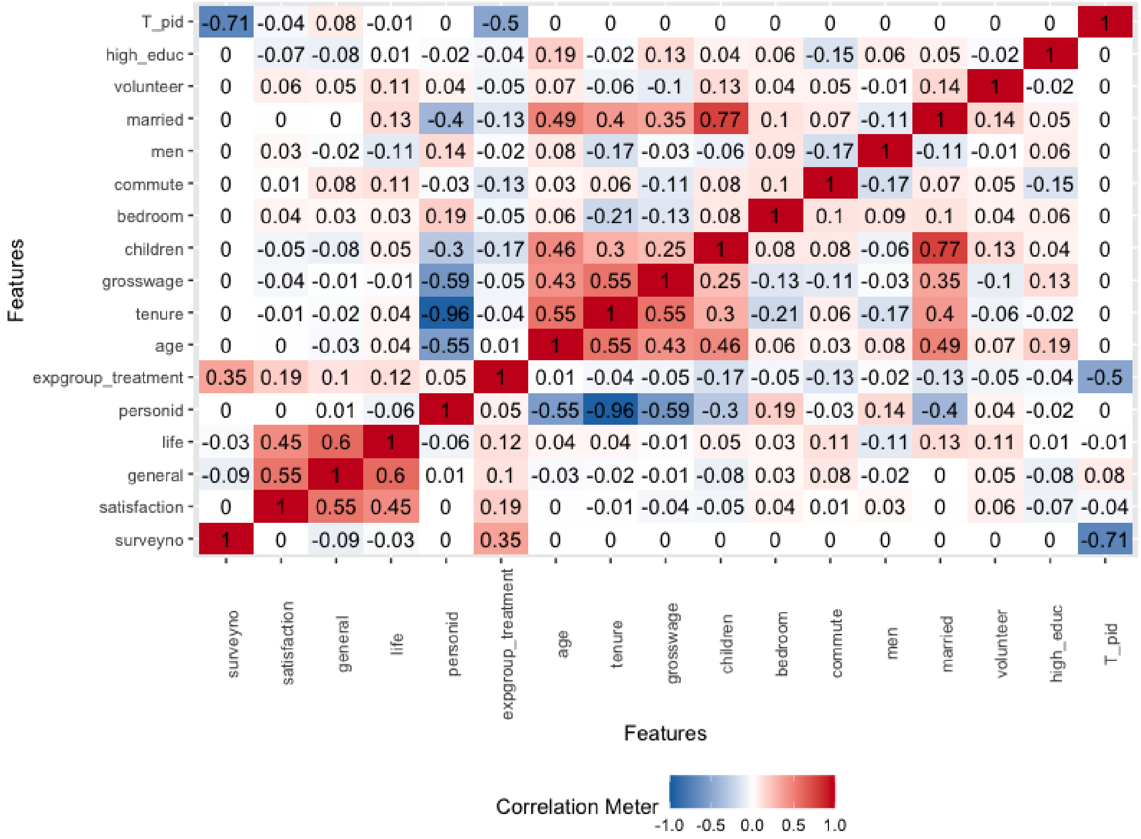


Figure 1: Correlation between variables



	surveyno	satisfaction	general	life	personid	expgroup_treatment	age	tenure	grosswage	children	bedroom	commute	men	married	volunteer	high_educ	T_pid
T_pid	-0.71	-0.04	0.08	-0.01	0	-0.5	0	0	0	0	0	0	0	0	0	0	1
high_educ	0	-0.07	-0.08	0.01	-0.02	-0.04	0.19	-0.02	0.13	0.04	0.06	-0.15	0.06	0.05	-0.02	1	0
volunteer	0	0.06	0.05	0.11	0.04	-0.05	0.07	-0.06	-0.1	0.13	0.04	0.05	-0.01	0.14	1	-0.02	0
married	0	0	0	0.13	-0.4	-0.13	0.49	0.4	0.35	0.77	0.1	0.07	-0.11	1	0.14	0.05	0
men	0	0.03	-0.02	-0.11	0.14	-0.02	0.08	-0.17	-0.03	-0.06	0.09	-0.17	1	-0.11	-0.01	0.06	0
commute	0	0.01	0.08	0.11	-0.03	-0.13	0.03	0.06	-0.11	0.08	0.1	1	-0.17	0.07	0.05	-0.15	0
bedroom	0	0.04	0.03	0.03	0.19	-0.05	0.06	-0.21	-0.13	0.08	1	0.1	0.09	0.1	0.04	0.06	0
children	0	-0.05	-0.08	0.05	-0.3	-0.17	0.46	0.3	0.25	1	0.08	0.08	-0.06	0.77	0.13	0.04	0
grosswage	0	-0.04	-0.01	-0.01	-0.59	-0.05	0.43	0.55	1	0.25	-0.13	-0.11	-0.03	0.35	-0.1	0.13	0
tenure	0	-0.01	-0.02	0.04	-0.96	-0.04	0.55	1	0.55	0.3	-0.21	0.06	-0.17	0.4	-0.06	-0.02	0
age	0	0	-0.03	0.04	-0.55	0.01	1	0.55	0.43	0.46	0.06	0.03	0.08	0.49	0.07	0.19	0
expgroup_treatment	0.35	0.19	0.1	0.12	0.05	1	0.01	-0.04	-0.05	-0.17	-0.05	-0.13	-0.02	-0.13	-0.05	-0.04	-0.5
personid	0	0	0.01	-0.06	1	0.05	-0.55	-0.96	-0.59	-0.3	0.19	-0.03	0.14	-0.4	0.04	-0.02	0
life	-0.03	0.45	0.6	1	-0.06	0.12	0.04	0.04	-0.01	0.05	0.03	0.11	-0.11	0.13	0.11	0.01	-0.01
general	-0.09	0.55	1	0.6	0.01	0.1	-0.03	-0.02	-0.01	-0.08	0.03	0.08	-0.02	0	0.05	-0.08	0.08
satisfaction	0	1	0.55	0.45	0	0.19	0	-0.01	-0.04	-0.05	0.04	0.01	0.03	0	0.06	-0.07	-0.04
surveyno	1	0	-0.09	-0.03	0	0.35	0	0	0	0	0	0	0	0	0	0	-0.71

Figure 1 displays the correlation between 17 variables, it shows that highest correlation is between married and children. Age and tenure also has 0.55 correlation among them.

Lowest correlation is between explanatory variables, high_educ and volunteer which -0.02 (Only considering correlation between explanatory variables which will be used in regression models)

Figure 2: Relation between age and tenure

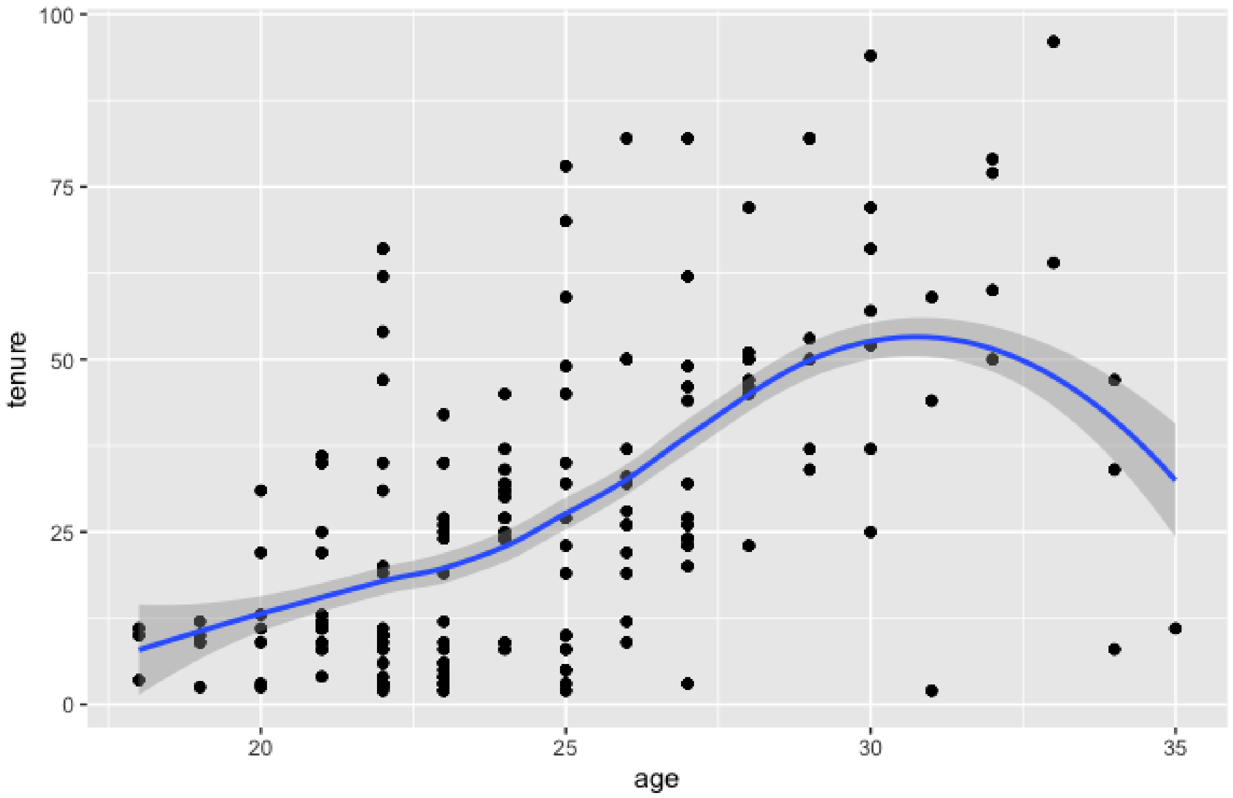


Table 6 showcases that there is slight positive relation between age and tenure, employees having higher age also tends to have higher tenure within the company, which means that they had join company at younger age and had stayed in company. So, this leads to 1st Hypothesis, where it can be said that such employees must be having high satisfaction levels in company, and they enjoy the work.

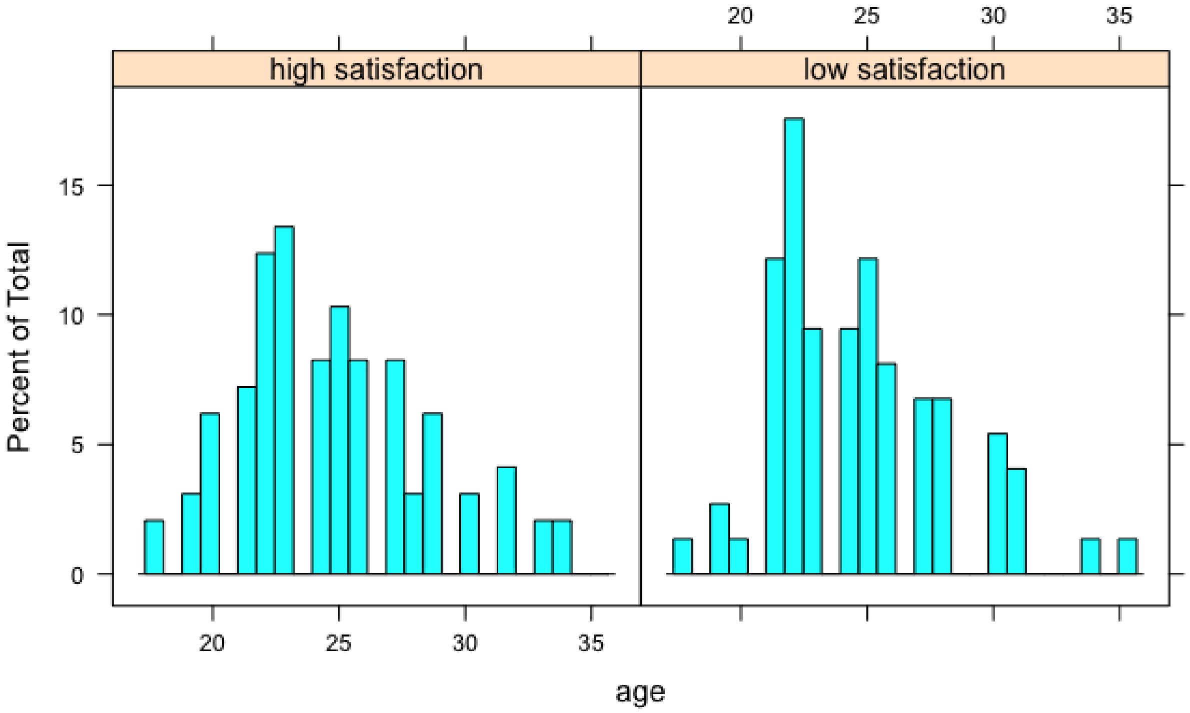
Data Analysis

Finally the model has been constructed with above discussed hypothesis and methods and following results are fetched and analyzed.

Before moving to other variables, dataset was divided into 2 parts using t_pid variable, as t_pid = 1 shows the start of the survey for each person, it means collected information is before treatment started.
t_pid = 0 states that information was collected during the treatment period for each person. There were total of 4 surveys done during treatment and 1 survey was conducted before treatment started.

Then the satisfaction level is divided as low satisfaction (with values = 1,2,3,4) and high satisfaction (having values = 5,6,7).
This has been done so that it can be analyzed that how much data lies with low satisfied employees and high satisfied employees.

Table 7 : Distribution of Satisfaction levels



Above figure displays that in younger age group level of satisfaction is lower as compared to high satisfaction.

Then lets divided the satisfaction level of complete Data set into 2 groups,
One before treatment and second one during treatment. It will give us a better idea of distribution of satisfaction level.

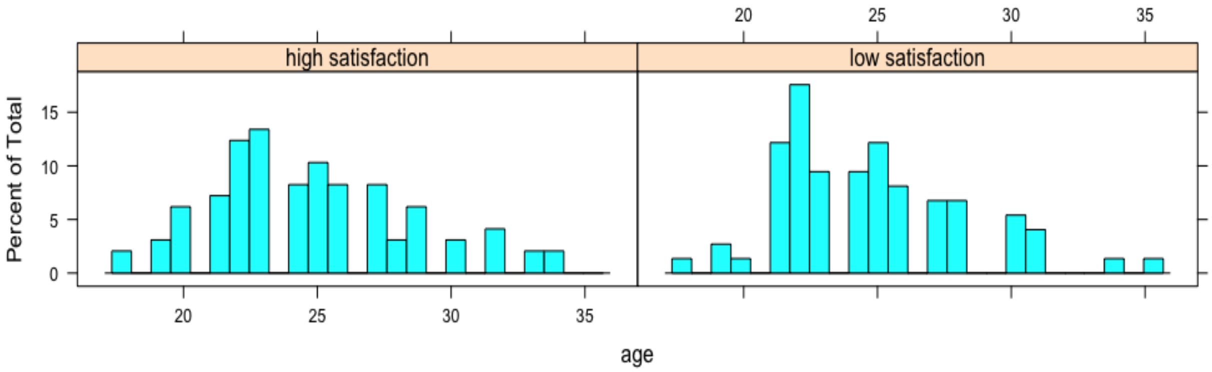


Table 7 : Distribution of Satisfaction level before treatment

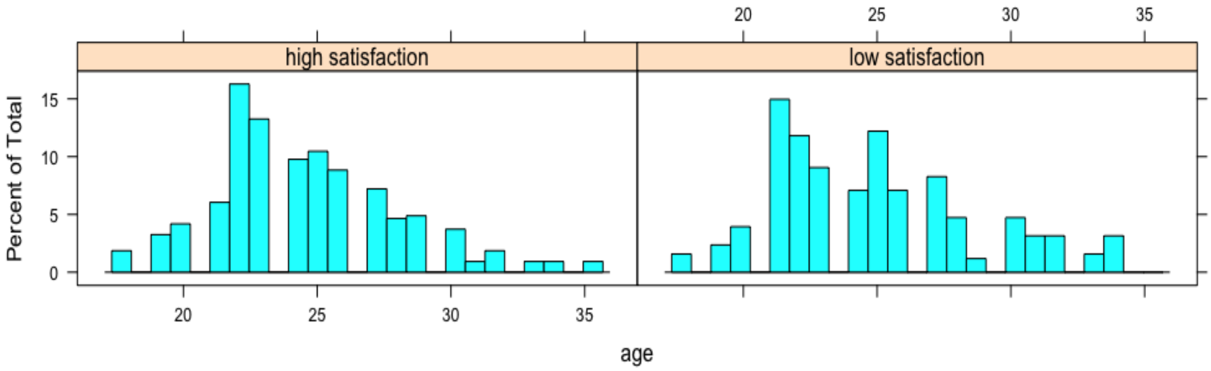


Table 7 : Distribution of Satisfaction level during treatment

Above figure states that before treatment started, younger age group between age 20 to 25 have low satisfaction levels, but once the treatment started, employees in same age group started to showcase higher satisfaction.

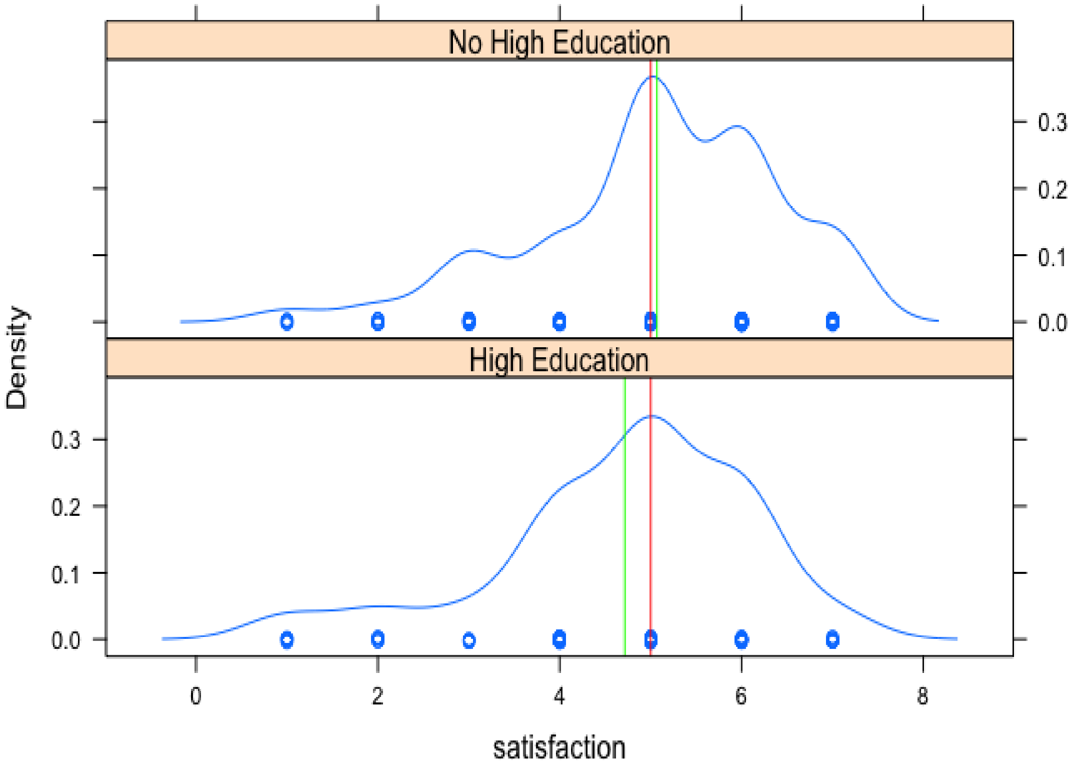
Let’s compute Density plot for Control Group: -

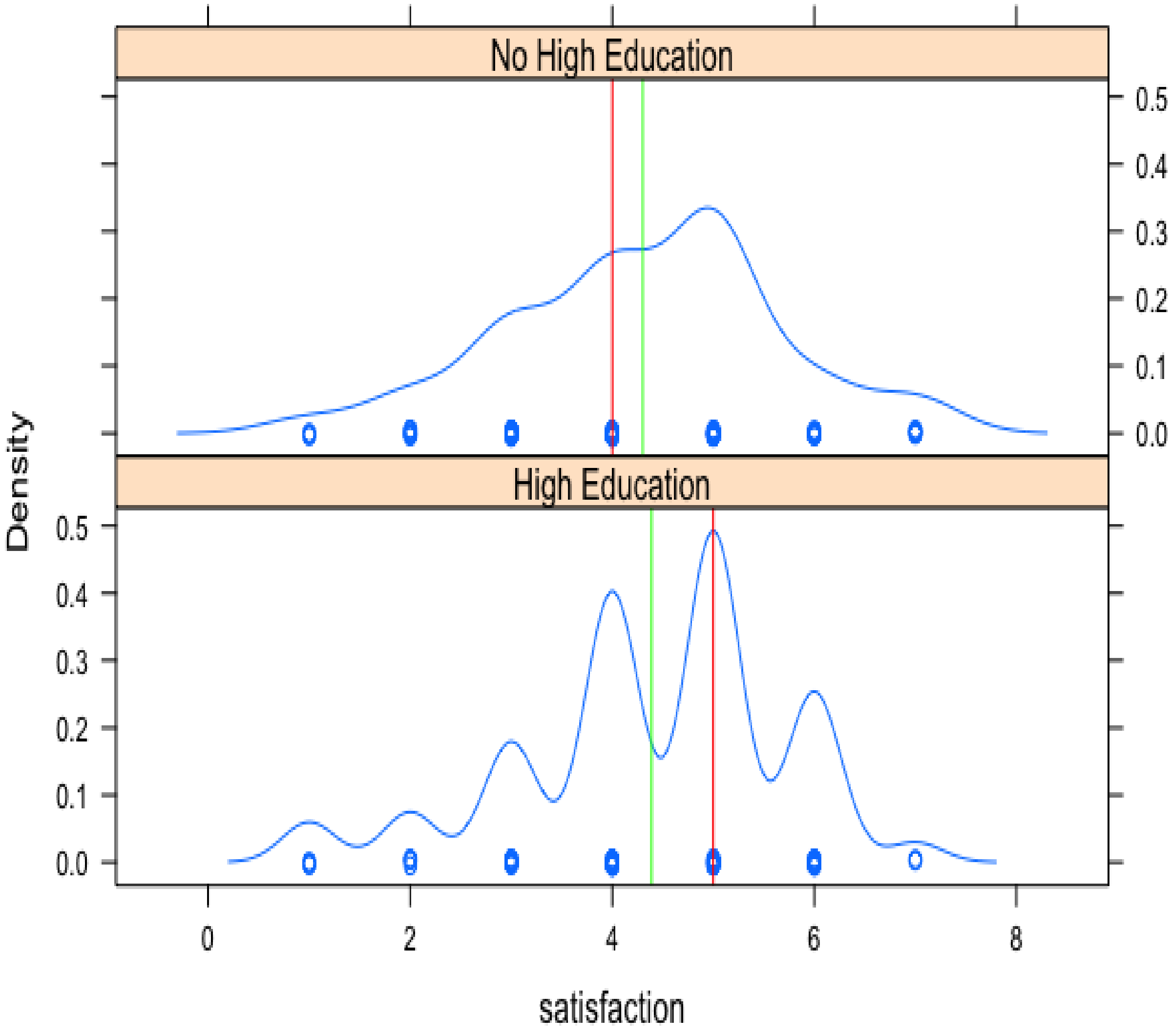
Below figure shows that, mean is almost same for Control group for employees having higher education vs employees not having high education. However, Median of employees with higher education is quite high compared to employees not having high education. This means that in the group of employees having higher education, significant number of employees show high levels of satisfaction which drags the median up.

Density plot for Treatment Group: -

Median values for treatment group are bit different, mean value remains the same. We are not able to derive the reasons behind why the behavior is such, so we need to run regression model to deduce this behavior.

Density plot for Control Group





Density plot for Treatment Group

```
Hypothesis 1 results,

===========================================================
                              Dependent variable:
                    ---------------------------------------
                                 satisfaction
                            (1)                 (2)
-----------------------------------------------------------
age                       -0.031              -0.001
                          (0.031)             (0.021)

tenure                   -0.048**             -0.005
                          (0.023)             (0.009)

Constant                 9.548***            5.397***
                          (2.105)             (0.929)

-----------------------------------------------------------
Observations                256                 428
R2                         0.027               0.002
Adjusted R2                0.012              -0.007
Residual Std. Error  1.316 (df = 251)    1.368 (df = 423)
F Statistic         1.758 (df = 4; 251) 0.263 (df = 4; 423)
===========================================================
Note:                           *p<0.1; **p<0.05; ***p<0.01
```

1st Model is for Control Group and 2nd model is for Treatment group.
OLS Regression model shows that there is no significant values of coefficient age and tenure, and this applies to both Control and Treatment Group.

So, lets analyse the same scenario using logistic regression, to see whether there is an impact, for that we have divided Satisfaction into 2 different levels (lower satisfaction having values 1 to 4 and higher satisfaction having values between 5 to 7)

```
Logistic regression output

==================================================
                           Dependent variable:
                       ---------------------------
                              satlevelcode
--------------------------------------------------
age                               0.021
                                 (0.039)

expgroup_treatment               2.421**
                                 (1.136)

tenure                           -0.022*
                                 (0.013)

age:expgroup_treatment           -0.061
                                 (0.046)

Constant                          1.675
                                 (1.502)

--------------------------------------------------
Observations                       684
Log Likelihood                  -430.950
Akaike Inf. Crit.                875.899
==================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

Age does not have significant impact on satisfaction level, also from interaction between age and expgroup_treatment we see that difference of effect of age on satisfaction between control and treatment group is not significantly different.

Hypothesis 2 results ——————-

1st model is for Control Group and 2nd model results are for Treatment Group.

```
=============================================================
                               Dependent variable:
                    -----------------------------------------
                                  satisfaction
                            (1)                  (2)
-------------------------------------------------------------
married                   -0.113                0.003
                          (0.200)              (0.175)

volunteer                 -0.246               0.409**
                          (0.289)              (0.195)

high_educ                  0.081              -0.342**
                          (0.172)              (0.139)

Constant                 5.042***             4.772***
                          (0.476)              (0.314)

-------------------------------------------------------------
Observations                256                  428
R2                         0.012                0.028
Adjusted R2               -0.008                0.016
Residual Std. Error  1.329 (df = 250)     1.352 (df = 422)
F Statistic         0.594 (df = 5; 250) 2.430** (df = 5; 422)
=============================================================
Note:                             *p<0.1; **p<0.05; ***p<0.01
```

It can be inferred that in Control Group, Volunteer and high_educ are not having any significant impact on Satisfaction, but in Treatment group, volunteer variable has a positive significance, which implies that employees who are working from home, which also indulges themselves in volunteer work tends to have higher satisfaction levels. However high_educ starts to show negative significance level, implying that employee having higher education are generally not satisfied. These results are in favour of our 2nd Hypothesis.

Conclusion

In Conclusion, companies are now following different hybrid models allowing employees to work from home, as studies have shown that people who work from home generally report higher satisfaction levels, and this will have an impact on employee's productivity.

We hope that above results help Human resources department in figuring a way out that all the employees get an option to work from home.

For Shanghai Ctrip call centre, results show that high age and tenure not necessarily means that employees will have high satisfaction.
For future studies, a thorough study can be conducted when all the employees are offered work from home in different hybrid models. Employees can do WFH on rotational basis and then it can showcase whether all employees show higher satisfaction levels. Also, we need to see that why Employees from treatment group having high education, have a negative impact on Satisfaction.

Appendix

```
library(tidyverse)
library(haven)
library(foreign)
library(stargazer)
library(ggplot2)
library(Hmisc)
library(chron)
library(lattice)
library(dummies)
library(lfe)
library(sandwich)
library(lmtest)
library(miceadds)
library(multiwayvcov)
library(margins)
library(foreign)
library(corrplot)
library(DataExplorer)
library(gridExtra)
install.packages("ggplot2")
library(funModeling)

setwd("~/Downloads")

#########################Importing the DataSet##############################
InitialData <- read_dta("Working from home-20210519/Satisfaction.dta")
view(InitialData)
write.csv(InitialData, file = "satisfaction.csv")


###################DataFrame for complete data###############################
SatisfactionData<-data.frame(InitialData)
x<-stargazer(SatisfactionData,type="text",title =,align=TRUE)
write.table(x, file = "results.txt", sep = ",", row.names = F,
            quote = FALSE)

#########################DataSet Exploration###############################
summary(SatisfactionData)
stargazer(SatisfactionData,type="text", align=TRUE,title = "Data Summary")
```

```
plot_correlation(SatisfactionData)
plot(SatisfactionData$satisfaction,xlab ="Person ID", ylab = "Satisfaction",
main= "Table 2 : Satisfaction by Person ID",col="red")
hist(SatisfactionData$age,xlab="Age", ylab="Number of Employee", labels =
TRUE,
     main = "Table 3 : Age Distribution in Company", col="red", ylim =
c(0,150),breaks = 24)
hist(SatisfactionData$tenure,xlab="Tenure", ylab="Number of Employee", labels
= TRUE,
     main = "Table 4 : Tenure Distribution in Company", col="red", ylim =
c(0,150),breaks = 24)

plot(SatisfactionData$age,SatisfactionData$tenure,  pch = 19, col = "red")
ggplot(SatisfactionData, aes(x=age, y=tenure)) + geom_point() +
  geom_smooth(method="auto", se=TRUE, fullrange=FALSE, level=0.95)




###############Setting up levels for Satisfaction###############
SatisfactionData<-within(SatisfactionData,{satislevel<-
ifelse(satisfaction<=4,"low satisfaction","high satisfaction")})
SatisfactionData<-within(SatisfactionData,{satlevelcode<-
ifelse(satislevel=="low satisfaction",0,1)})

###################### Distribution of Satisfaction levels ################
a <- histogram(~age | satislevel, data=ControlData1,nint=25, main=" Table 7 :
Distribution of Satisfaction level before treatment")
b <- histogram(~age | satislevel, data=TreatmentData1,nint=25, main=" Table 7
: Distribution of Satisfaction level during treatment")
grid.arrange(a,b,nrow=2)

###############Setting up levels for High education and Volunteer############
SatisfactionData<-within(SatisfactionData,{highedu<-ifelse(high_educ == 0,"No
High Education","High Education")})
SatisfactionData<-within(SatisfactionData,{vol<-ifelse(volunteer == 0,"NO
Volunteer","Do Volunteer")})
a <- histogram(~ volunteer | high_educ, data=ControlData1,nint=25, main="
Table 7 : Distribution of Satisfaction level before treatment")
b <- histogram(~ volunteer | high_educ, data=TreatmentData1,nint=25, main="
Table 7 : Distribution of Satisfaction level during treatment")
grid.arrange(a,b,nrow=2)

######DataSet for Control Group when T_pid =1 (Before Treatment Started)#####
ControlData1 <- SatisfactionData[SatisfactionData$T_pid == "1", ]
view(ControlData1)
ControlData <- SatisfactionData[SatisfactionData$expgroup_treatment == "0", ]
```

```
view(ControlData)
plot(ControlData)


####DataSet for Treatment Group and Control Group with T_pid=0 (During
Treatment Period)#####

TreatmentData1 <- SatisfactionData[SatisfactionData$T_pid == "0", ]
view(TreatmentData1)
FinalTreatmentData <- TreatmentData1[TreatmentData1$expgroup_treatment ==
"1", ]
FinalControlData <- TreatmentData1[TreatmentData1$expgroup_treatment == "0",
]
view(FinalTreatmentData)

###########Density Plot for Comparing Treatment and Control Data############

densityplot(~ satisfaction | highedu, data=FinalTreatmentData, layout=c(1,2),
            panel=function(x,...)
            {
              panel.densityplot(x,...)
              panel.abline(v=quantile(x,.5),col.line = "red")
              panel.abline(v=mean(x), col.line = "green")
            }
)

densityplot(~ satisfaction | highedu, data=FinalControlData, layout=c(1,2),
            panel=function(x,...)
            {
              panel.densityplot(x,...)
              panel.abline(v=quantile(x,.5),col.line = "red")
              panel.abline(v=mean(x), col.line = "green")
            }
)




####################### Model Design for Hypothesis 1 #######################
lm1 <- lm(satisfaction ~ age + tenure + personid + surveyno, data =
FinalControlData)
lm2 <- lm(satisfaction ~ age + tenure + personid + surveyno, data =
FinalTreatmentData)
lm3 <- lm(satisfaction ~ age*tenure + married + commute + volunteer +
personid + surveyno, data = FinalTreatmentData)
lm3 <- lm(satisfaction ~ age + tenure + grosswage, data = FinalTreatmentData)
```

```
lm4 <- lm(satisfaction ~ age + tenure + grosswage + married, data =
FinalTreatmentData)
lm5 <- lm(satisfaction ~ age + tenure  + married*children, data =
FinalTreatmentData)

lm10 <- glm(satlevelcode ~ age*expgroup_treatment + tenure+ personid +
surveyno,
            data = TreatmentData1,family=binomial(link="logit"))

stargazer(lm4,lm5,lm6,type="text", align=TRUE,title = "Table 7 :OLS
model")

y <- stargazer(lm1,lm2, type="text", omit =c("surveyno","personid"))
write.table(y, file = "regression.txt", sep = ",", row.names = F,
            quote = FALSE)
z <- stargazer(lm10, type="text", omit =c("surveyno","personid"))
write.table(z, file = "logit.txt", sep = ",", row.names = F,
            quote = FALSE)

#################### Model Design for Hypothesis 2 ##########################
lm6 <- lm(satisfaction ~ married + volunteer + high_educ + personid +
surveyno , data = FinalTreatmentData)
lm8 <- lm(satisfaction ~ married + volunteer + high_educ + personid +
surveyno, data = FinalControlData)


stargazer(lm8,lm6, type="text", omit =c("surveyno","personid"))
```

References

Bodin, C. B. (2008). Office Types in Relation to Health, Well-Being and Job Satisfaction among employees.

Liang, N. B. (2014). DOES WORKING FROM HOME WORK? EVIDENCE FROM CHINESE EXPERIMENT.

Mateyka, P. J. (2010). Home-Based Workers in United States

.