T-Test

- A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related to certain features. It is mostly used when the data sets follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

- There are three types of t-test:

1. One sample t-test
2. Unpaired/independent t-test
   a. Student's t-test
   b. Welch's t-test
3. Paired/dependent t-test

- All the hypotheses henceforth have been conducted and validated using Welch's two sample independent t-test.

- **Formula** (Welch's test) **:-**

$$t = \frac{(x_1 - x_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

- Tool used :- **RStudio** for R Programming

T-TEST for DATASET - 01

Importing all the necessary libraries with the dataset :-

```
library(dplyr)
library(tidyverse)
library(sm)
library(corrplot)
ds1 <- read.csv(file.choose(), header = T)
View(ds1)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone se
2	5575-GNVDE	Male	0	No	No	34	Yes	No
3	3668-QPYBK	Male	0	No	No	2	Yes	No
4	7795-CFOCW	Male	0	No	No	45	No	No phone se
5	9237-HQITU	Female	0	No	No	2	Yes	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes
7	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes
8	6713-OKOMC	Female	0	No	No	10	No	No phone se
9	7892-POOKP	Female	0	Yes	No	28	Yes	Yes
10	6388-TABGU	Male	0	No	Yes	62	Yes	No

Showing 1 to 13 of 7,043 entries, 21 total columns

Converting the categorical variables into numeric values :-

```
cat<-data.matrix(ds1)
df_cat <- data.frame(cat)
View(df_cat)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
1	5376	1	0	2	1	1	1	2
2	3963	2	0	1	1	34	2	1
3	2565	2	0	1	1	2	2	1
4	5536	2	0	1	1	45	1	2
5	6512	1	0	1	1	2	2	1
6	6552	1	0	1	1	8	2	3
7	1003	2	0	1	2	22	2	3
8	4771	1	0	1	1	10	1	2
9	5605	1	0	2	1	28	2	3
10	4535	2	0	1	2	62	2	1
11	6872	2	0	2	2	13	2	1

Showing 1 to 12 of 7,043 entries, 21 total columns

Viewing the structure of the dataset :-

```
str(df_cat)
```

```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : num  5376 3963 2565 5536 6512 ...
 $ gender          : num  1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : num  2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : num  1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : num  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : num  1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : num  2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : num  1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : num  1 3 3 3 1 1 1 3 1 3 ...
 $ OnlineBackup    : num  3 1 3 1 1 1 3 1 1 3 ...
 $ DeviceProtection: num  1 3 1 3 1 3 1 1 3 1 ...
 $ TechSupport     : num  1 1 1 3 1 1 1 1 3 1 ...
 $ StreamingTV     : num  1 1 1 1 1 3 3 1 3 1 ...
 $ StreamingMovies : num  1 1 1 1 1 3 1 1 3 1 ...
 $ Contract        : num  1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: num  2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : num  3 4 4 1 3 3 2 4 3 1 ...
```

Summary statistics for dataset :-

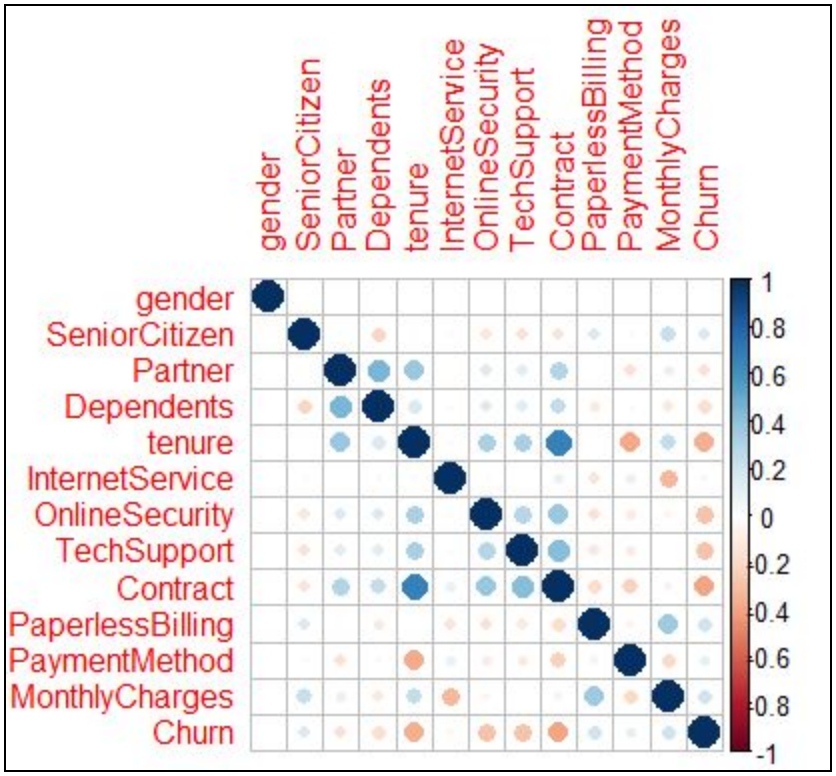
```
summary(df_cat)
```

```
   customerID       gender      SeniorCitizen       Partner      
 Min.   :   1   Min.   :1.000   Min.   :0.0000   Min.   :1.000  
 1st Qu.:1762   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000  
 Median :3522   Median :2.000   Median :0.0000   Median :1.000  
 Mean   :3522   Mean   :1.505   Mean   :0.1621   Mean   :1.483  
 3rd Qu.:5282   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:2.000  
 Max.   :7043   Max.   :2.000   Max.   :1.0000   Max.   :2.000  

   Dependents      tenure       PhoneService   MultipleLines  
 Min.   :1.0   Min.   : 0.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.0   1st Qu.: 9.00   1st Qu.:2.000   1st Qu.:1.000  
 Median :1.0   Median :29.00   Median :2.000   Median :2.000  
 Mean   :1.3   Mean   :32.37   Mean   :1.903   Mean   :1.941  
 3rd Qu.:2.0   3rd Qu.:55.00   3rd Qu.:2.000   3rd Qu.:3.000  
 Max.   :2.0   Max.   :72.00   Max.   :2.000   Max.   :3.000  
```

Correlation plot :-

```
M <- cor(df_cat[,c(2,3,4,5,6,9,10,13,16,17,18,19,21)])
corrplot(M, method = 'circle')
```



Some correlations that can be used for t-test :-

1.tenure-online security

2.tenure-contract

3.tenure-churn (negative correlation)

→ Based on these correlations, a few hypotheses can be formulated.

→ The confidence interval is 95 % which means that the l.o.s is 5%. Therefore, $\alpha = 0.05$ (consequently, $\alpha/2 = 0.025$ for two tailed t-test)

1]tenure-online security :- [NOTE:- Online security: 1 = 'No', 2 = 'No internet', 3 = 'Yes']

Fig-1.1: *Boxplot for tenure-online security*

```
ggplot(df_cat,aes(factor(OnlineSecurity),tenure))+
geom_boxplot(aes(colour = factor(OnlineSecurity)))+ ggtitle('Measure
of Tenure as per Online Security')+ coord_flip()
```

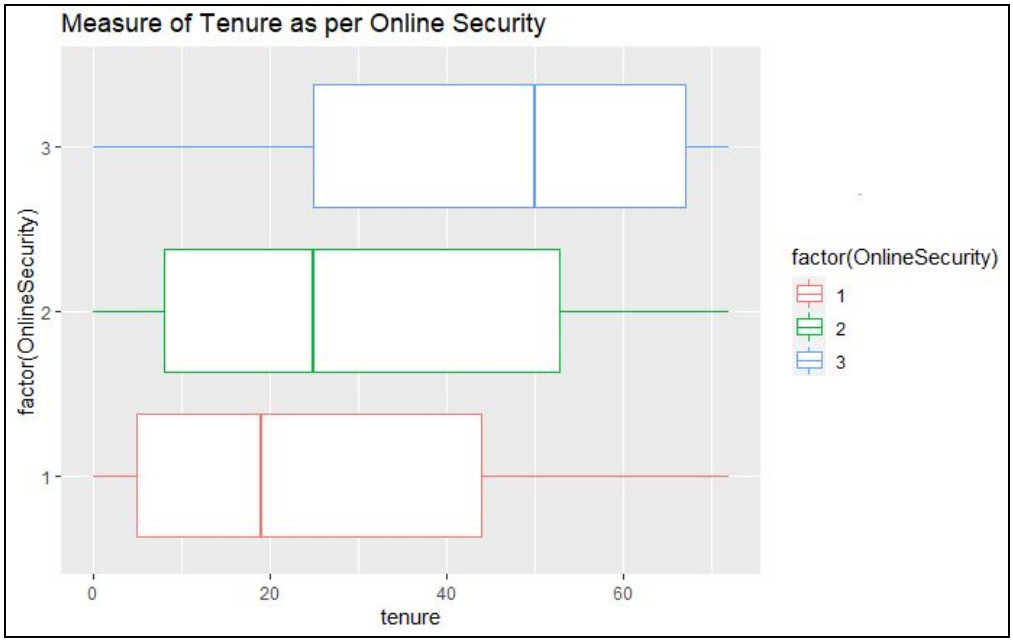
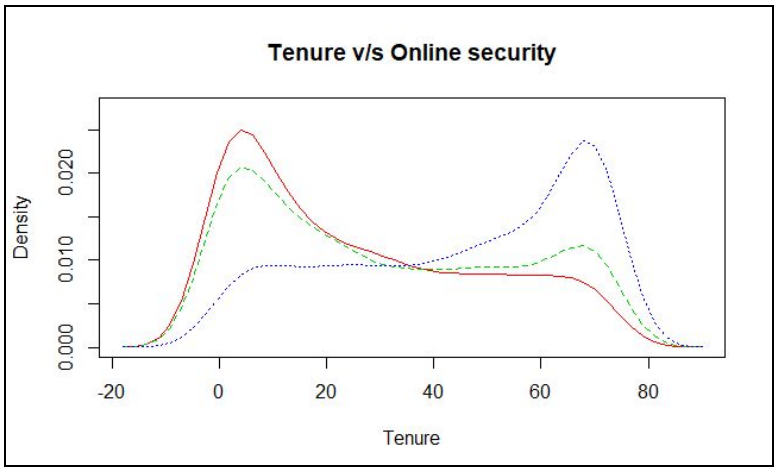


Fig-1.2: *Comparative density plots for tenure-online security.*

```
sm.density.compare(df_cat$tenure, df_cat$OnlineSecurity,xlab=
'Tenure')
title(main = 'Tenure v/s Online security')
```



Comment: From the boxplots and comparative density plots it can be concluded that the distribution for online security = 'yes' is negatively skewed. For online security = 'no', the distribution is positively skewed and as per the boxplot, the distribution is somewhat positively skewed for 'No-internet'.

HYPOTHESIS:-

H_0: Online security does not affect tenure.
H_1: Online security affects tenure.

T-test:

```
y.rows = df_cat[df_cat$OnlineSecurity == 3,]
n.rows = df_cat[df_cat$OnlineSecurity == 1,]
t.test(y.rows$tenure, n.rows$tenure)
```

```
> y.rows = df_cat[df_cat$OnlineSecurity == 3,]
> n.rows = df_cat[df_cat$OnlineSecurity == 1,]
> t.test(y.rows$tenure, n.rows$tenure)

        Welch Two Sample t-test

data:  y.rows$tenure and n.rows$tenure
t = 29.921, df = 4119.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 17.93799 20.45359
sample estimates:
mean of x mean of y
 45.04656  25.85077
```

Conclusion: Since p-value = 2.2e-16 << 0.05, we **reject** H_0 and accept H_1 which states that online security affects the tenure of customers .

2]tenure-contract :- [NOTE:- Contract:1 = 'Month-to-month', 2 = 'One-year', 3 = 'Two-year']

Fig-2.1: *Boxplots for tenure-contract*

```
ggplot(df_cat,aes(factor(Contract),tenure))+ geom_boxplot(aes(colour
= factor(Contract)))+ ggtitle('Measure of Tenure as per Contract')+
coord_flip()
```

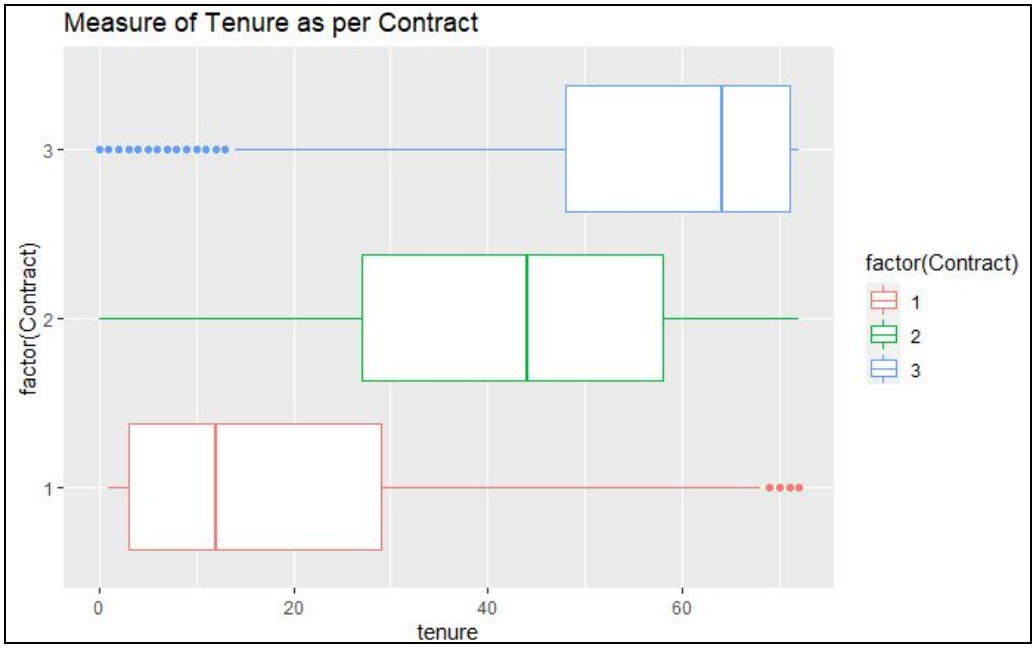
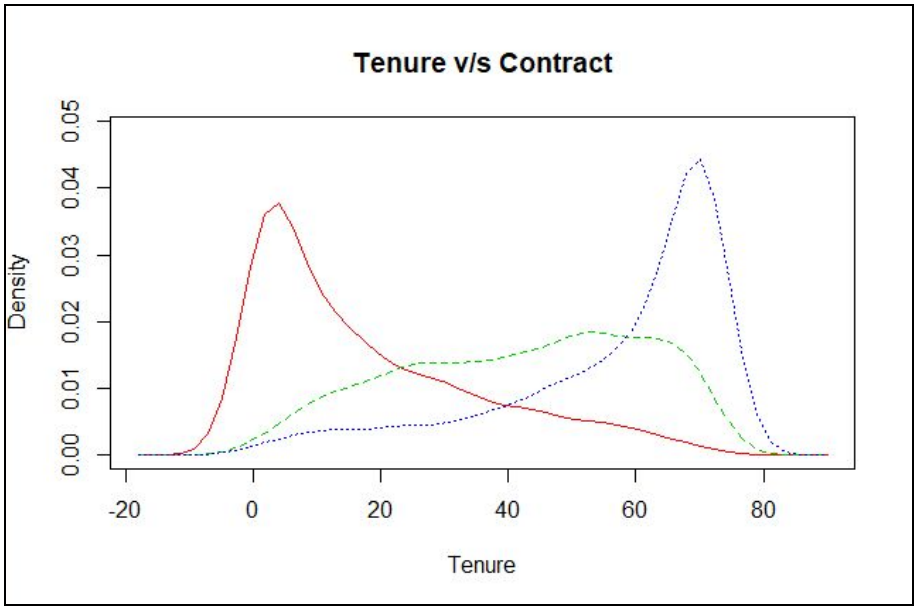


Fig-2.2: *Comparative density plots for tenure-contract.*

```
sm.density.compare(df_cat$tenure, df_cat$Contract,xlab= 'Tenure')
title(main = 'Tenure v/s Contract')
```



Comment: The distribution for 'two-year' contract is highly negatively skewed. The 'one-year' contract shows a slight negative skew whereas the 'month-to-month' contract is positively skewed.

HYPOTHESIS:-

H_0 : The type of contract does not affect tenure.
H_1 : The type of contract affects the tenure.

T-test:

```
o.rows = df_cat[df_cat$Contract == 2,]
t.rows = df_cat[df_cat$Contract == 3,]
t.test(o.rows$tenure, t.rows$tenure)
```

```
> o.rows = df_cat[df_cat$Contract == 2,]
> t.rows = df_cat[df_cat$Contract == 3,]
> t.test(o.rows$tenure, t.rows$tenure)

        Welch Two Sample t-test

data:  o.rows$tenure and t.rows$tenure
t = -22.106, df = 3061.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.99331 -13.38729
sample estimates:
mean of x mean of y
 42.04481  56.73510
```

Conclusion: Since p-value = 2.2e-16 << 0.05, we **reject** H_0 and accept H_1. Thus, it can be concluded that the type of contract affects the tenure.

3]tenure-churn :- [NOTE:- Churn: 1 = 'No', 2 = 'Yes']

Fig-3.1: *Boxplots for tenure-churn*

```
ggplot(df_cat,aes(factor(Churn),tenure))+geom_boxplot(aes(colour =
factor(Churn)))+ggtitle('Measure of Tenure as per Churn')+
coord_flip()
```

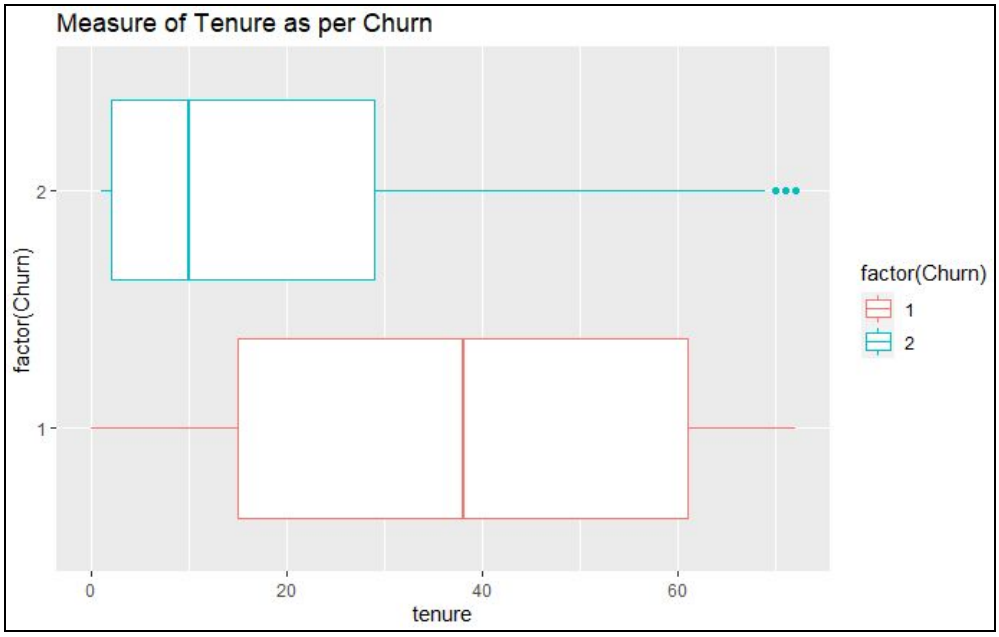
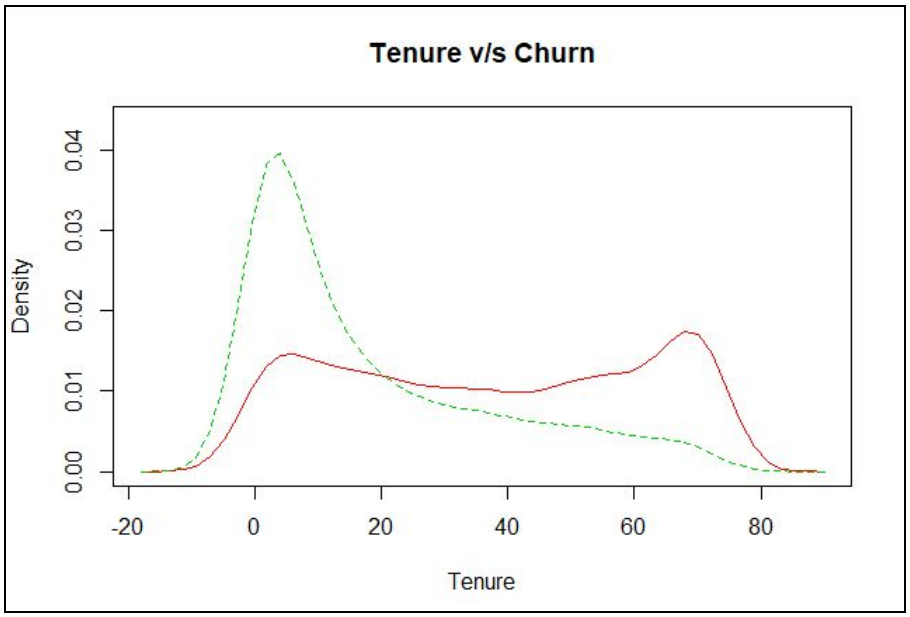


Fig-3.2: *Comparative density plots for tenure-churn.*

```
sm.density.compare(df_cat$tenure, df_cat$Churn,xlab= 'Tenure')
title(main = 'Tenure v/s Churn')
```



Comment: The distribution for churn = 'no' is symmetrically skewed (acc. to the boxplot) whereas the distribution for churn = 'yes' is highly positively skewed.

HYPOTHESIS:-

H_0 : Churning does not affect tenure.
H_1 : Churning affects tenure.

T-test:

```
yes.rows = df_cat[df_cat$Churn == 2,]
no.rows = df_cat[df_cat$Churn == 1,]
t.test(yes.rows$tenure, no.rows$tenure)
```

```
> yes.rows = df_cat[df_cat$Churn == 2,]
> no.rows = df_cat[df_cat$Churn == 1,]
> t.test(yes.rows$tenure, no.rows$tenure)

        Welch Two Sample t-test

data:  yes.rows$tenure and no.rows$tenure
t = -34.824, df = 4048.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.69378 -18.48789
sample estimates:
mean of x mean of y
 17.97913  37.56997
```

Conclusion: Since p-value = 2.2e-16 << 0.05, we **reject** H_0 and accept H_1 which states that churning affects the tenure of customers.

T-TEST for DATASET - 02

Importing all the necessary libraries with the dataset :-

```
library(dplyr)
library(tidyverse)
library(sm)
library(corrplot)
ds2 <- read.csv(file.choose(), header = T)
View(ds2)
```

	state	account.length	area.code	phone.number	international.plan	voice.mail.plan	number.vmail.messa
1	KS	128	415	382-4657	no	yes	25
2	OH	107	415	371-7191	no	yes	26
3	NJ	137	415	358-1921	no	no	0
4	OH	84	408	375-9999	yes	no	0
5	OK	75	415	330-6626	yes	no	0
6	AL	118	510	391-8027	yes	no	0
7	MA	121	510	355-9993	no	yes	24
8	MO	147	415	329-9001	yes	no	0
9	LA	117	408	335-4719	no	no	0
10	WV	141	415	330-8173	yes	yes	37
11	IN	65	415	329-6603	no	no	0

Showing 1 to 12 of 3,333 entries, 21 total columns

Converting the categorical variables into numeric values :-

```
cat<-data.matrix(ds2)
df_cat <- data.frame(cat)
View(df_cat)
```

	state	account.length	area.code	phone.number	international.plan	voice.mail.plan	number.vmail.messa
1	17	128	415	1927	1	2	25
2	36	107	415	1576	1	2	26
3	32	137	415	1118	1	1	0
4	36	84	408	1708	2	1	0
5	37	75	415	111	2	1	0
6	2	118	510	2254	2	1	0
7	20	121	510	1048	1	2	24
8	25	147	415	81	2	1	0
9	19	117	408	292	1	1	0
10	50	141	415	118	2	2	37
11	16	65	415	71	1	1	0

Showing 1 to 12 of 3,333 entries, 21 total columns

Viewing the structure of the dataset :-

```
str(df_cat)
```

```
'data.frame':   3333 obs. of  21 variables:
 $ state                 : num  17 36 32 36 37 2 20 25 19 50 ...
 $ account.length        : num  128 107 137 84 75 118 121 147 117 141 ...
 $ area.code             : num  415 415 415 408 415 510 510 415 408 415 ...
 $ phone.number          : num  1927 1576 1118 1708 111 ...
 $ international.plan    : num  1 1 1 2 2 2 1 2 1 2 ...
 $ voice.mail.plan       : num  2 2 1 1 1 1 2 1 1 2 ...
 $ number.vmail.messages : num  25 26 0 0 0 0 24 0 0 37 ...
 $ total.day.minutes     : num  265 162 243 299 167 ...
 $ total.day.calls       : num  110 123 114 71 113 98 88 79 97 84 ...
 $ total.day.charge      : num  45.1 27.5 41.4 50.9 28.3 ...
 $ total.eve.minutes     : num  197.4 195.5 121.2 61.9 148.3 ...
 $ total.eve.calls       : num  99 103 110 88 122 101 108 94 80 111 ...
 $ total.eve.charge      : num  16.78 16.62 10.3 5.26 12.61 ...
 $ total.night.minutes   : num  245 254 163 197 187 ...
 $ total.night.calls     : num  91 103 104 89 121 118 118 96 90 97 ...
 $ total.night.charge    : num  11.01 11.45 7.32 8.86 8.41 ...
 $ total.intl.minutes    : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2
 ...
 $ total.intl.calls      : num  3 3 5 7 3 6 7 6 4 5 ...
 $ total.intl.charge     : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.0
2 ...
 $ customer.service.calls: num  1 1 0 2 3 0 3 0 1 0 ...
 $ churn                 : num  1 1 1 1 1 1 1 1 1 1 ...
```

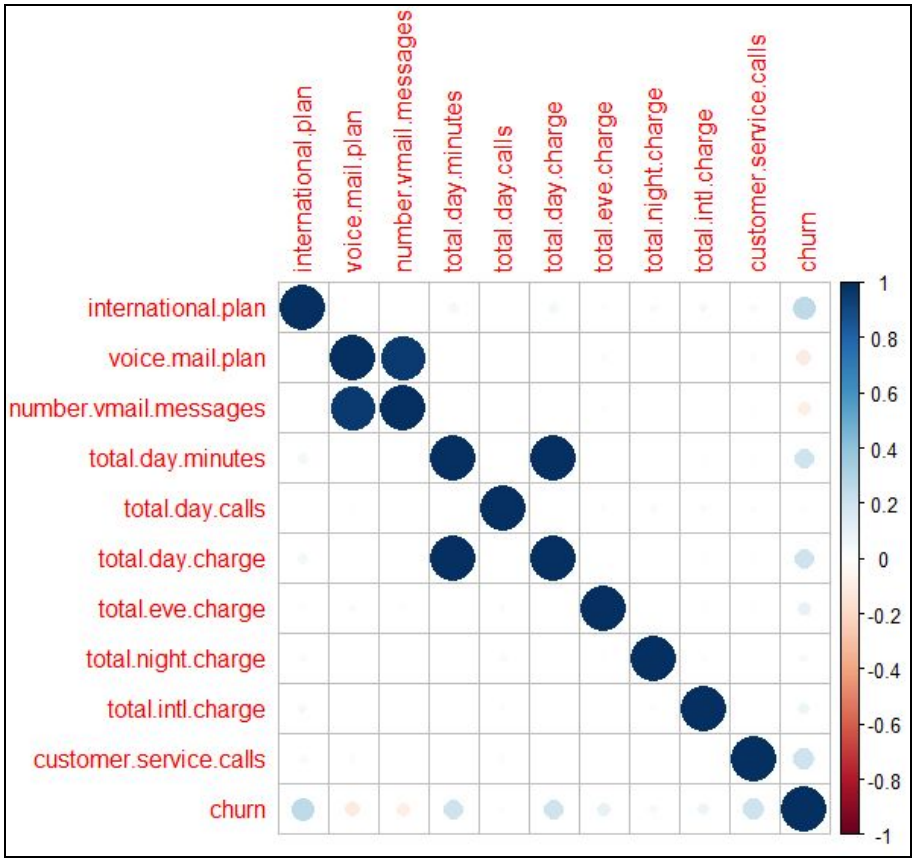
Summary statistics for dataset :-

```
summary(df_cat)
```

```
     state       account.length    area.code      phone.number
 Min.   : 1.00   Min.   :  1.0   Min.   :408.0   Min.   :   1
 1st Qu.:15.00   1st Qu.: 74.0   1st Qu.:408.0   1st Qu.: 834
 Median :27.00   Median :101.0   Median :415.0   Median :1667
 Mean   :27.06   Mean   :101.1   Mean   :437.2   Mean   :1667
 3rd Qu.:40.00   3rd Qu.:127.0   3rd Qu.:510.0   3rd Qu.:2500
 Max.   :51.00   Max.   :243.0   Max.   :510.0   Max.   :3333
 international.plan voice.mail.plan number.vmail.messages
 Min.   :1.000      Min.   :1.000   Min.   : 0.000
 1st Qu.:1.000      1st Qu.:1.000   1st Qu.: 0.000
 Median :1.000      Median :1.000   Median : 0.000
 Mean   :1.097      Mean   :1.277   Mean   : 8.099
 3rd Qu.:1.000      3rd Qu.:2.000   3rd Qu.:20.000
 Max.   :2.000      Max.   :2.000   Max.   :51.000
 total.day.minutes total.day.calls total.day.charge total.eve.minutes
 Min.   :  0.0     Min.   :  0.0   Min.   : 0.00    Min.   :  0.0
 1st Qu.:143.7     1st Qu.: 87.0   1st Qu.:24.43    1st Qu.:166.6
 Median :179.4     Median :101.0   Median :30.50    Median :201.4
 Mean   :179.8     Mean   :100.4   Mean   :30.56    Mean   :201.0
 3rd Qu.:216.4     3rd Qu.:114.0   3rd Qu.:36.79    3rd Qu.:235.3
 Max.   :350.8     Max.   :165.0   Max.   :59.64    Max.   :363.7
 total.eve.calls total.eve.charge total.night.minutes total.night.calls
 Min.   :  0.0   Min.   : 0.00    Min.   : 23.2       Min.   : 33.0
 1st Qu.: 87.0   1st Qu.:14.16    1st Qu.:167.0       1st Qu.: 87.0
 Median :100.0   Median :17.12    Median :201.2       Median :100.0
 Mean   :100.1   Mean   :17.08    Mean   :200.9       Mean   :100.1
 3rd Qu.:114.0   3rd Qu.:20.00    3rd Qu.:235.3       3rd Qu.:113.0
 Max.   :170.0   Max.   :30.91    Max.   :395.0       Max.   :175.0
```

Correlation plot :-

```
M <- cor(df_cat[,c(2,3,4,5,6,9,10,13,16,17,18,19,21)])
corrplot(M, method = 'circle')
```



Some correlations that can be considered for t-test :-

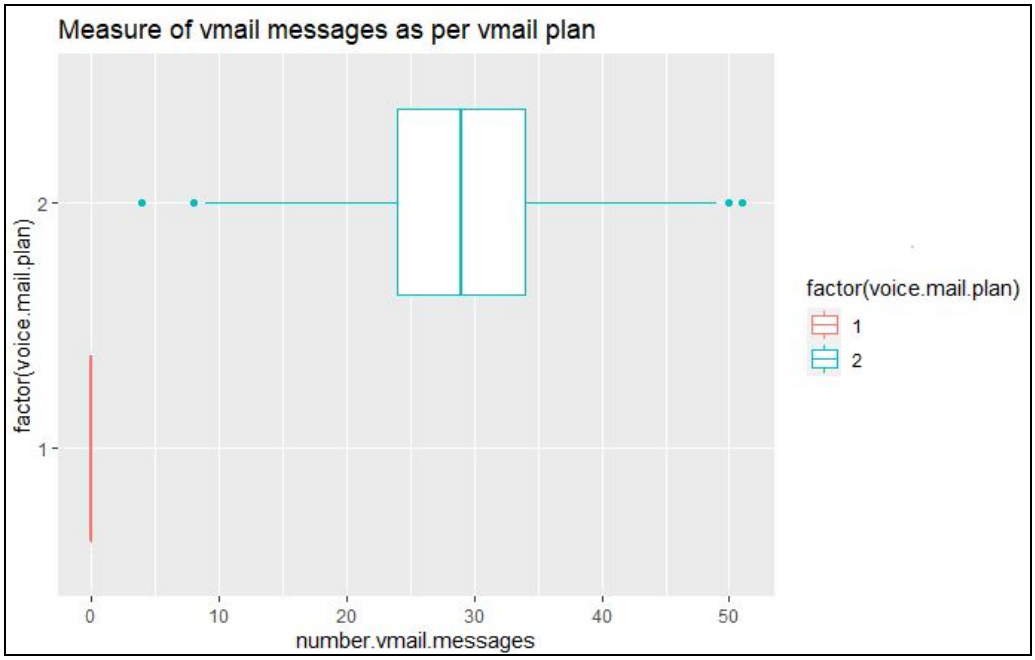
1. number.vmail.messages - voice.mail.plan
3. total.day.charge - churn
3. total.intl.charge - churn

→ A few hypotheses can be formulated. Confidence interval is 95 % which means that the l.o.s is 5%. Therefore, α = 0.05 (consequently, $\alpha/2$ = 0.025 for two tailed t-test)

1]number.vmail.messages-voice.mail.plan :- [NOTE:- Voicemail plan: 1 = 'No', 2 = 'Yes']

Fig-1.1: *Boxplot for number.vmail.messages-voice.mail.plan*

```
ggplot(df_cat, aes(factor(voice.mail.plan),number.vmail.messages))
+geom_boxplot(aes(colour = factor(voice.mail.plan)))+ggtitle('Measure
of vmail messages as per vmail plan')+ coord_flip()
```



Comment: The distribution for voicemail plan = 'No' is almost negligible whereas the distribution for voicemail plan = 'yes' is roughly symmetrically skewed.

Hypothesis →

HYPOTHESIS :-

H_0: Voicemail plan does not have an impact on the number of voicemail messages.
H_1: Voicemail plan has an impact on the number of voicemail messages.

T-test :

```
w.rows = df_cat[df_cat$voice.mail.plan == 2,]
wo.rows = df_cat[df_cat$voice.mail.plan == 1,]
t.test(w.rows$number.vmail.messages, wo.rows$number.vmail.messages)
```

```
> w.rows = df_cat[df_cat$voice.mail.plan == 2,]
> wo.rows = df_cat[df_cat$voice.mail.plan == 1,]
> t.test(w.rows$number.vmail.messages, wo.rows$number.vmail.messages)

        Welch Two Sample t-test

data:  w.rows$number.vmail.messages and wo.rows$number.vmail.messages
t = 117.61, df = 921, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 28.78910 29.76622
sample estimates:
mean of x mean of y
 29.27766   0.00000
```

Conclusion: Since p-value = 2.2e-16 << 0.05, we **reject** H_0 and accept H_1 which states that the subscription of a voicemail plan does in fact, have an impact on the number of voicemail messages.

2]total.day.charge - churn :- [NOTE:- Churn : 1 = 'False', 2 = 'True']

Fig-2.1: *Boxplot for total.day.charge and churn*

```
ggplot(df_cat,aes(factor(churn),total.day.charge))+ geom_boxplot(aes
(colour = factor(churn)))+ ggtitle('Measure of Total daily charge as
per Churn')+coord_flip()
```

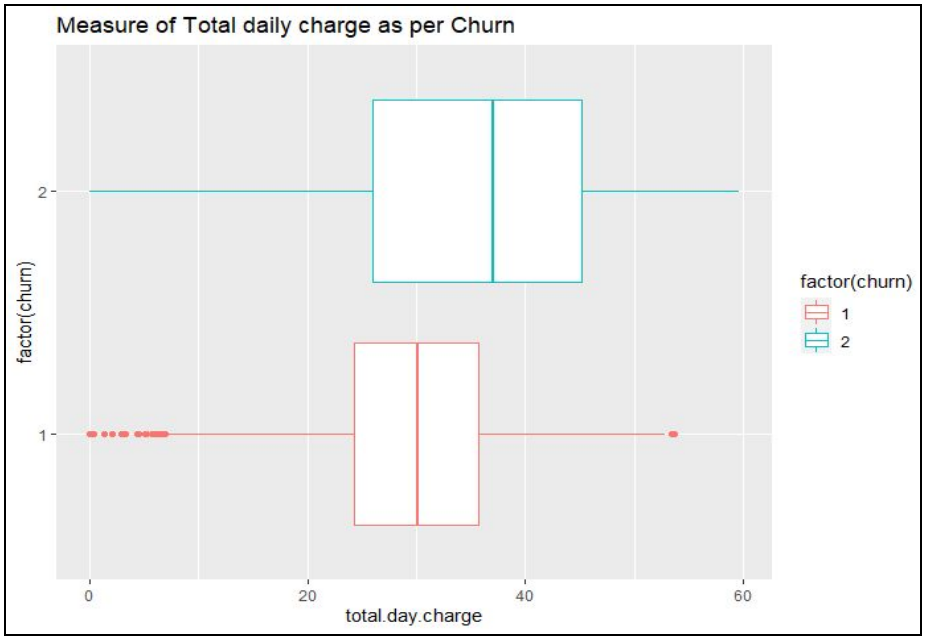
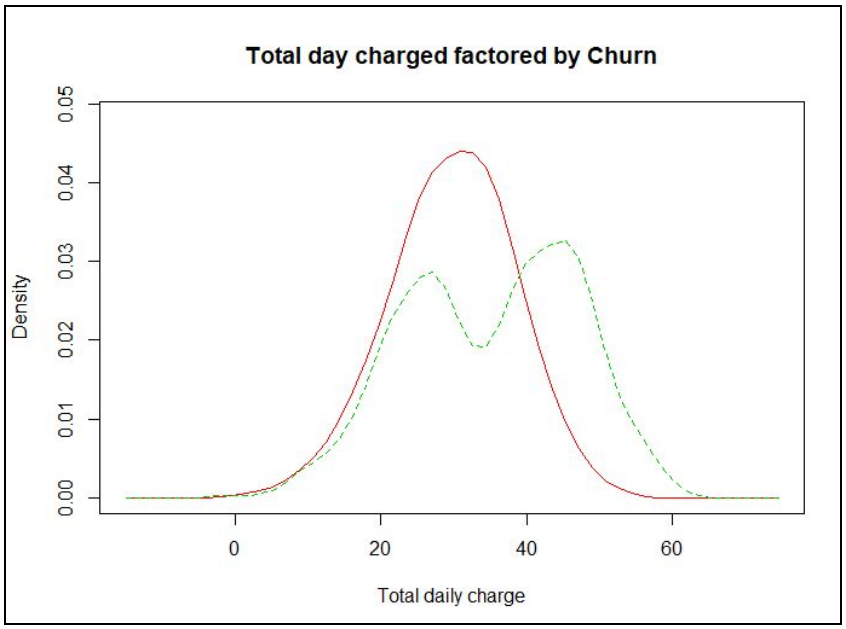


Fig-2.2: *Comparative density plot for total.day.charge and churn*

```
sm.density.compare(df_cat$total.day.charge, df_cat$churn, xlab =
'Total daily charge')
title(main = 'Total day charged factored by Churn')
```



Comment: The curve for churn = 'False' is symmetrically and normally distributed. Churn = 'True' shows a slight negative skew.

HYPOTHESIS :-

H_0: Churning is not associated with total day charge.
H_1: Churning is associated with total day charge.

T-test :

```
f.rows = df_cat[df_cat$churn == 1,]
t.rows = df_cat[df_cat$churn == 2,]
t.test(f.rows$total.day.charge, t.rows$total.day.charge)
```

```
> f.rows = df_cat[df_cat$churn == 1,]
> t.rows = df_cat[df_cat$churn == 2,]
> t.test(f.rows$total.day.charge, t.rows$total.day.charge)

        Welch Two Sample t-test

data:  f.rows$total.day.charge and t.rows$total.day.charge
t = -9.6845, df = 571.51, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.489770 -4.301231
sample estimates:
mean of x mean of y
 29.78042  35.17592
```

Conclusion: Since p-value = 2.2e-16 << 0.05, we **reject** H_0 and accept H_1. This would mean that churning is associated with total day charge.

3]total.intl.charge - churn :- [NOTE:- Churn : 1 = 'False', 2 = 'True']

Fig-3.1: *Boxplot for total.intl.charge and churn*

```
ggplot(df_cat, aes(factor(churn),
total.intl.charge))+geom_boxplot(aes(colour =
factor(churn)))+ggtitle('Measure of Total international charge as per
Churn')+coord_flip()
```

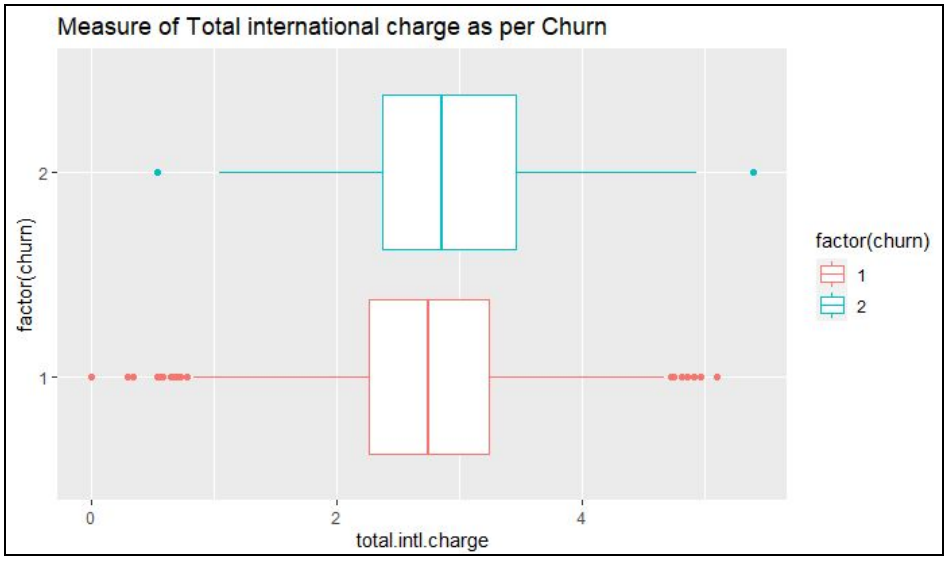
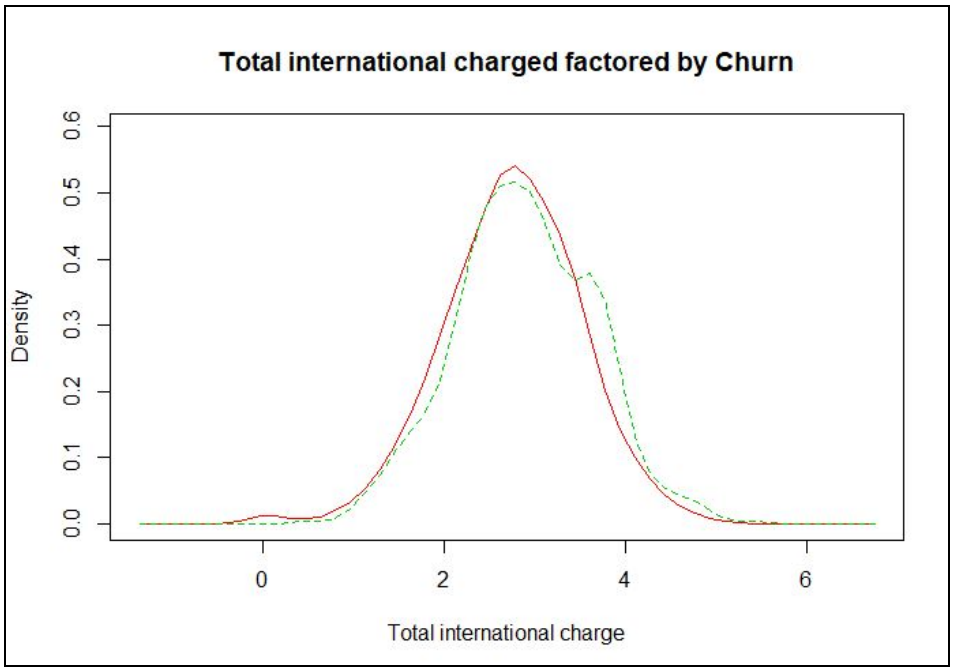


Fig-3.2: *Comparative density plot for total.intl.charge and churn*

```
sm.density.compare(df_cat$total.intl.charge, df_cat$churn, xlab =
'Total international charge')
title(main = 'Total international charged factored by Churn')
```



Comment: The curve for churn = 'False' is symmetrically and normally distributed. Churn = 'True' tends to show a positively skewed distribution.

HYPOTHESIS :-

H_0: Churning is not associated with international charge.
H_1: Churning is associated with international charges.

T-test :

```
t.rows = df_cat[df_cat$churn == 2,]
f.rows = df_cat[df_cat$churn == 1,]
t.test(t.rows$total.intl.charge, f.rows$total.intl.charge)
```

```
> t.rows = df_cat[df_cat$churn == 2,]
> f.rows = df_cat[df_cat$churn == 1,]
> t.test(t.rows$total.intl.charge, f.rows$total.intl.charge)

        Welch Two Sample t-test

data:  t.rows$total.intl.charge and f.rows$total.intl.charge
t = 3.9399, df = 654.88, p-value = 9.026e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.07330691 0.21897510
sample estimates:
mean of x mean of y
 2.889545  2.743404
```

Conclusion: Since p-value = 0.00009 << 0.05, we **reject** H_0 and accept H_1. It can be concluded that churning is associated with international charges.
