DR. ALVIN'S PUBLICATIONS

STATISTICS WITH R

BY DR. ALVIN ANG

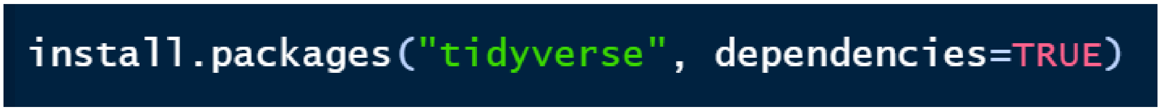
TABLE OF CONTENTS

I. TIDYVERSE PACKAGE

Tons of great Data Wrangling with R here:

https://www.marsja.se/how-to-rename-column-or-columns-in-r-with-dplyr/

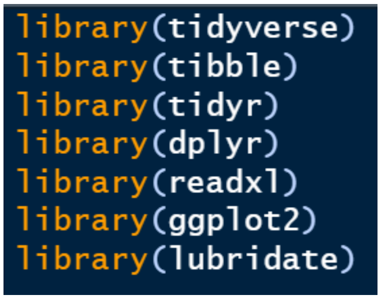
Install Tidyverse Package to R:



```
install.packages("tidyverse", dependencies=TRUE)
```

- install.packages("tidyverse", dependencies=TRUE)

Run the following libraries:



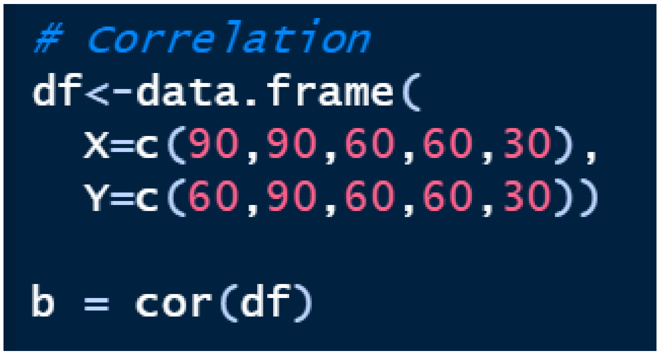
```
library(tidyverse)
library(tibble)
library(tidyr)
library(dplyr)
library(readxl)
library(ggplot2)
library(lubridate)
```

- library(tidyverse)
- library(tibble)
- library(tidyr)
- library(dplyr)
- library(readxl)
- library(ggplot2)
- library(lubridate)

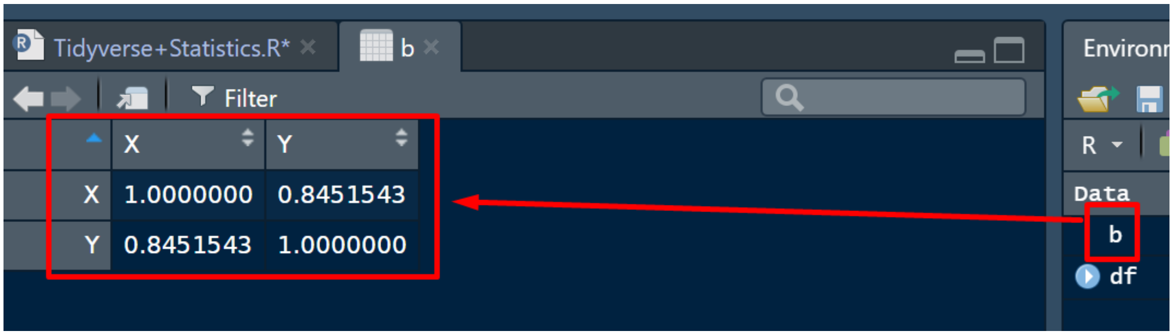
II. TIDYVERSE STATISTICS

File can be found here: https://www.alvinang.sg/s/Statistics-with-Tidyverse-by-Dr-Alvin-Ang.R

A. CORRELATION I

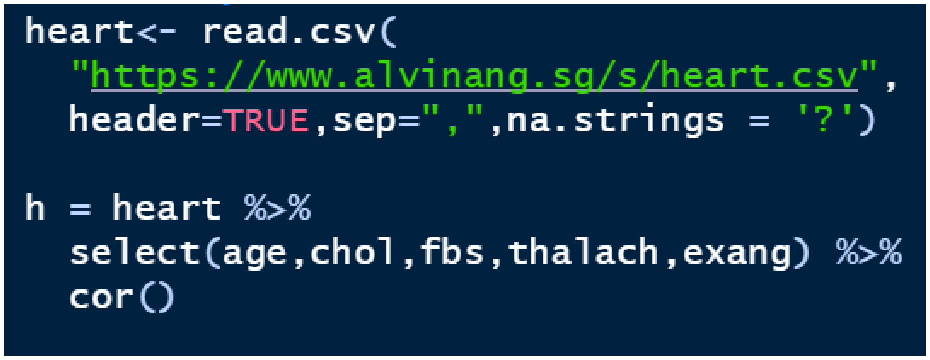


```
# Correlation
df<-data.frame(
  X=c(90,90,60,60,30),
  Y=c(60,90,60,60,30))

b = cor(df)
```

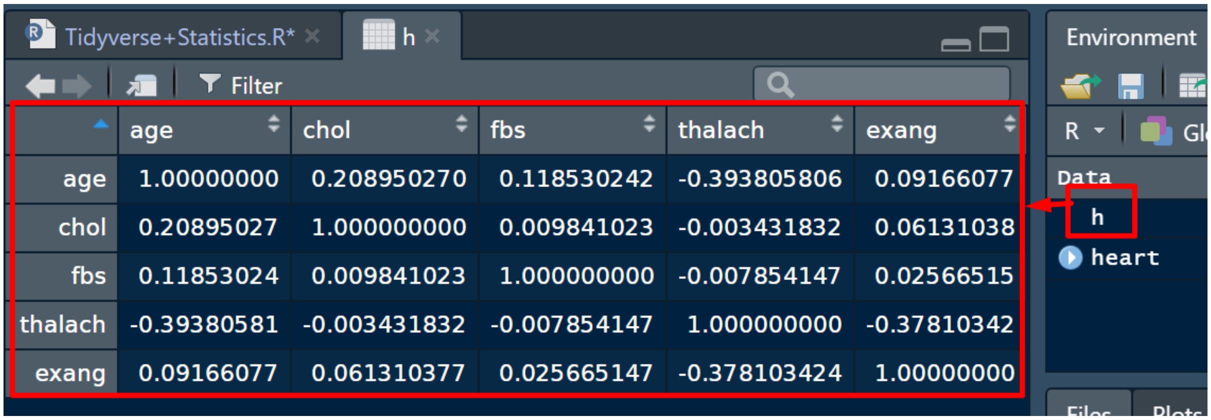


```
          X         Y
X 1.0000000 0.8451543
Y 0.8451543 1.0000000
```

B. CORRELATION II



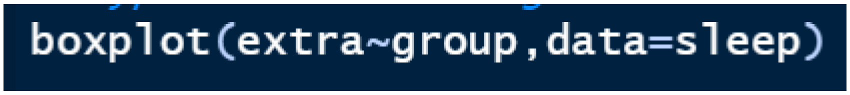
```
heart<- read.csv(
  "https://www.alvinang.sg/s/heart.csv",
  header=TRUE,sep=",",na.strings = '?')

h = heart %>%
  select(age,chol,fbs,thalach,exang) %>%
  cor()
```



	age	chol	fbs	thalach	exang
age	1.00000000	0.208950270	0.118530242	-0.393805806	0.09166077
chol	0.20895027	1.000000000	0.009841023	-0.003431832	0.06131038
fbs	0.11853024	0.009841023	1.000000000	-0.007854147	0.02566515
thalach	-0.39380581	-0.003431832	-0.007854147	1.000000000	-0.37810342
exang	0.09166077	0.061310377	0.025665147	-0.378103424	1.00000000

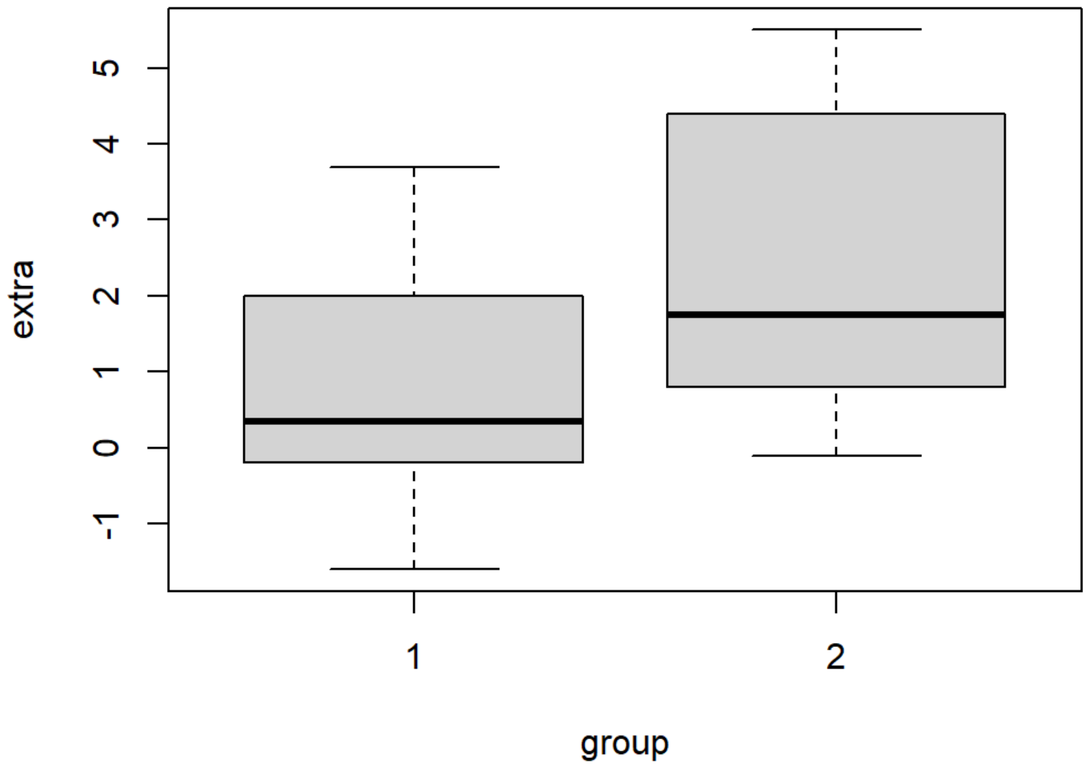
C. HYPOTHESIS TESTING (TWO TAILED TEST)

1. BOXPLOT

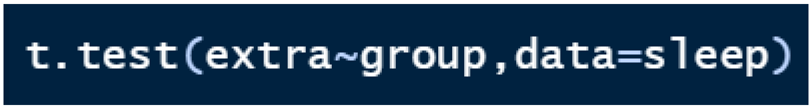


```
boxplot(extra~group,data=sleep)
```

doesn't seem like there's a significant difference between the 2 groups.....due to overlap....

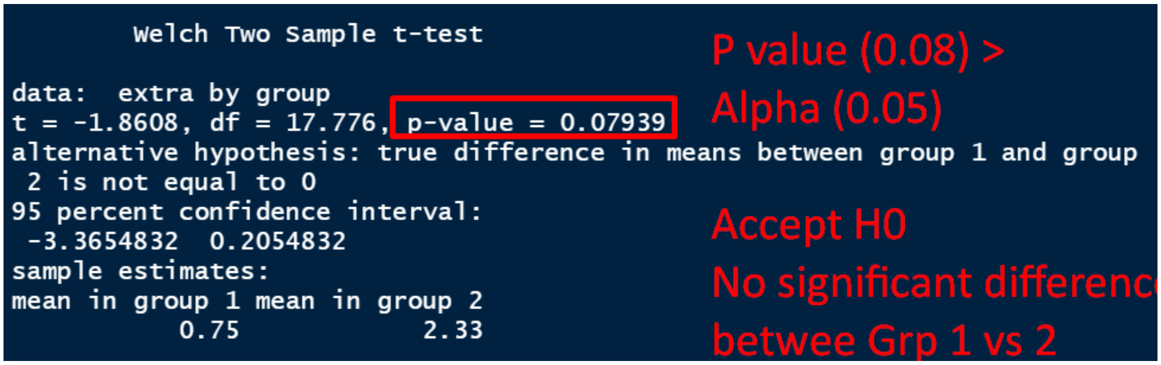


2. TWO SAMPLE T TEST



```
t.test(extra~group,data=sleep)
```

1. Stating the Claim → 2 Tailed test:
    a. H0: Mean sleep of Grp 1 = Mean sleep of Grp 2
    b. H1: Mean sleep of Grp 1 ≠ Mean sleep of Grp 2
2. Running the Test



```
        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means between group 1 and group
 2 is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2 
           0.75            2.33 
```

P value (0.08) > Alpha (0.05)

Accept H0

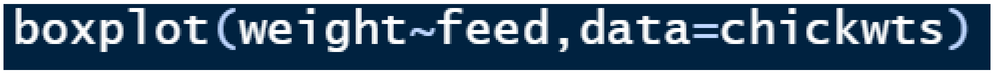
No significant differenc

betwee Grp 1 vs 2

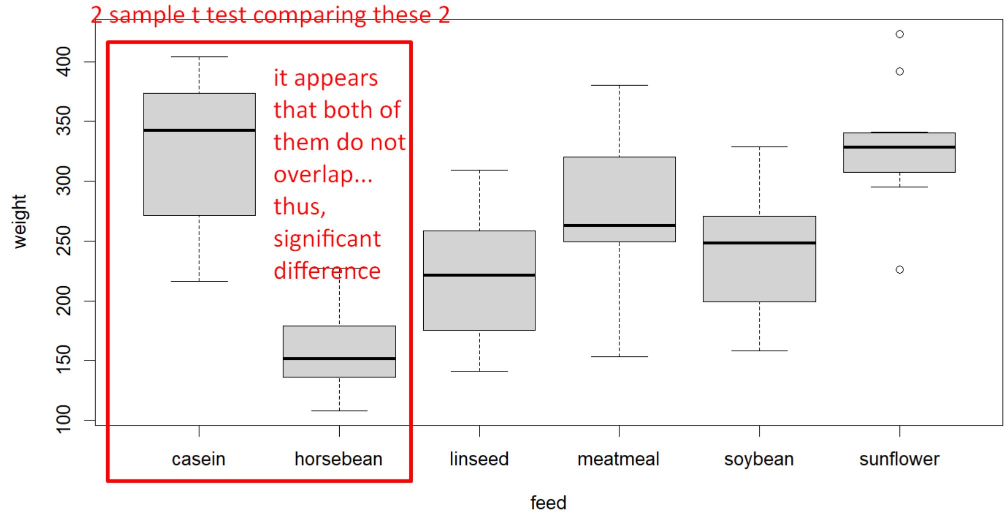
3. Conclusion:
    a. We accept H1
    b. There's NO significant difference between the Sleep amounts of Grp 1 vs Grp 2.

D. HYPOTHESIS TESTING (ONE TAILED TEST)

1. BOX PLOT



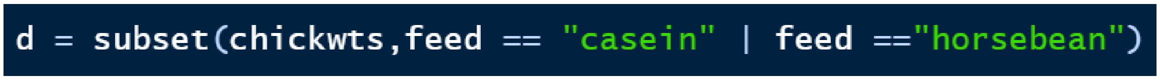
```
boxplot(weight~feed,data=chickwts)
```



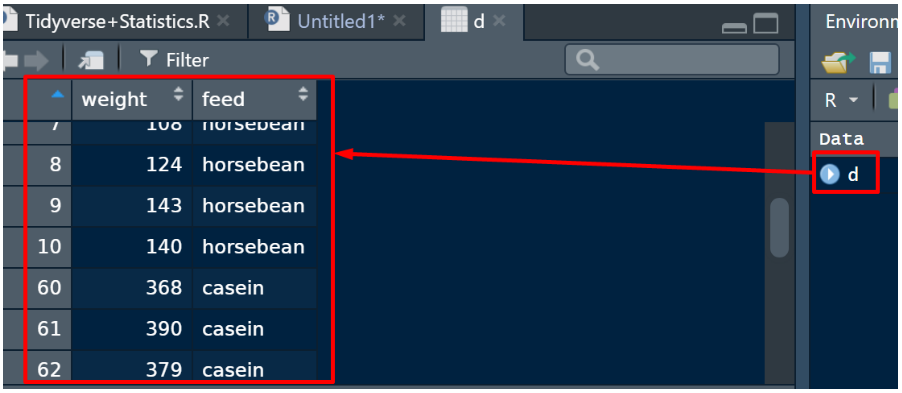
2 sample t test comparing these 2

it appears that both of them do not overlap... thus, significant difference

2. SELECTING COLUMNS

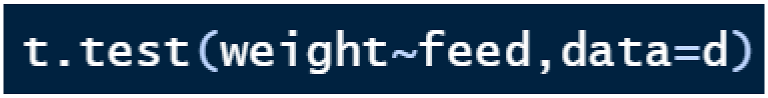


```
d = subset(chickwts,feed == "casein" | feed =="horsebean")
```



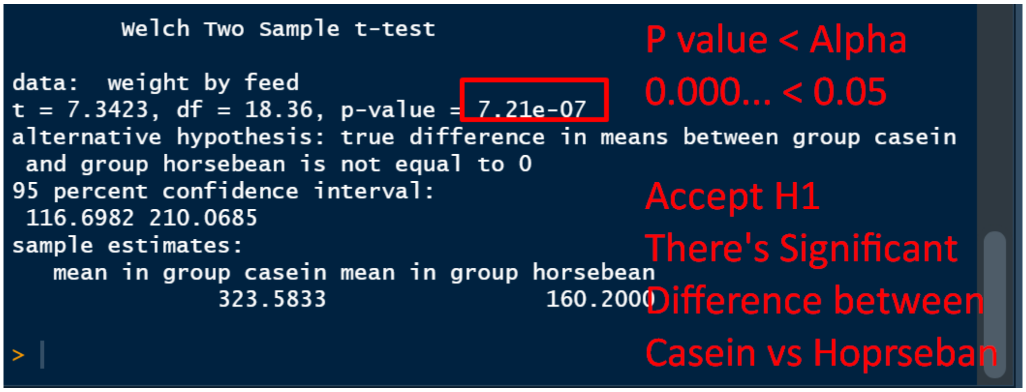
	weight	feed
7	108	horsebean
8	124	horsebean
9	143	horsebean
10	140	horsebean
60	368	casein
61	390	casein
62	379	casein

3. 2 TAILED T TEST



```
t.test(weight~feed,data=d)
```

- H0: Casein = Horsebean
- H1: Casein ≠ Horsebean



```
        Welch Two Sample t-test

data:  weight by feed
t = 7.3423, df = 18.36, p-value = 7.21e-07
alternative hypothesis: true difference in means between group casein
 and group horsebean is not equal to 0
95 percent confidence interval:
 116.6982 210.0685
sample estimates:
   mean in group casein mean in group horsebean
               323.5833                160.2000

> |
```

P value < Alpha

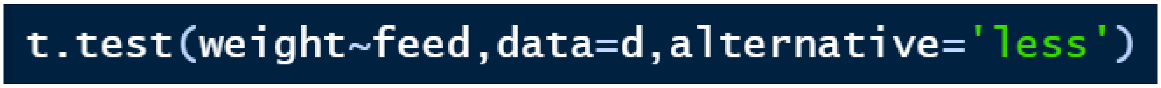
0.000... < 0.05

Accept H1

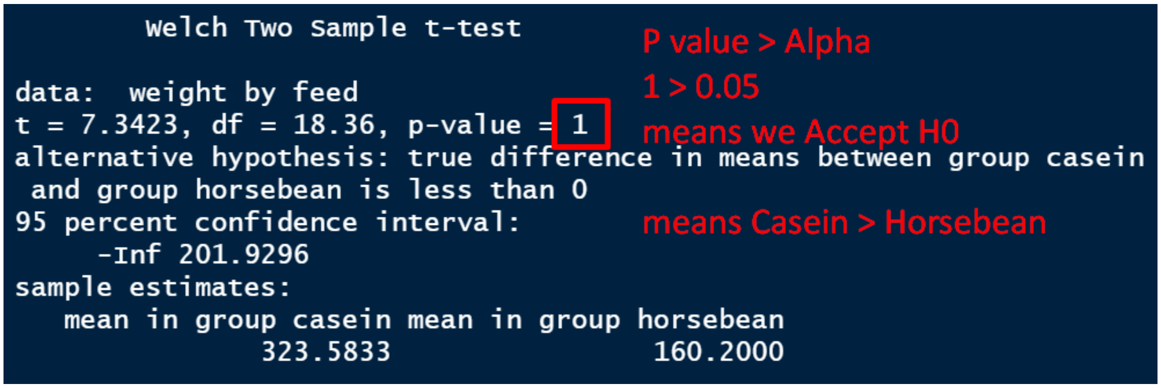
There's Significant Difference between Casein vs Hoprseban

4. 1 TAILED T TEST

- H0: Casein >= Horsebean
- H1: Casein < Horsebean



```
t.test(weight~feed,data=d,alternative='less')
```



```
        Welch Two Sample t-test

data:  weight by feed
t = 7.3423, df = 18.36, p-value = 1
alternative hypothesis: true difference in means between group casein
 and group horsebean is less than 0
95 percent confidence interval:
     -Inf 201.9296
sample estimates:
   mean in group casein mean in group horsebean
               323.5833                160.2000
```

P value > Alpha

1 > 0.05

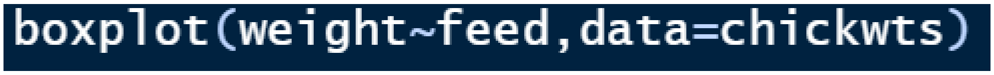
means we Accept H0

means Casein > Horsebean

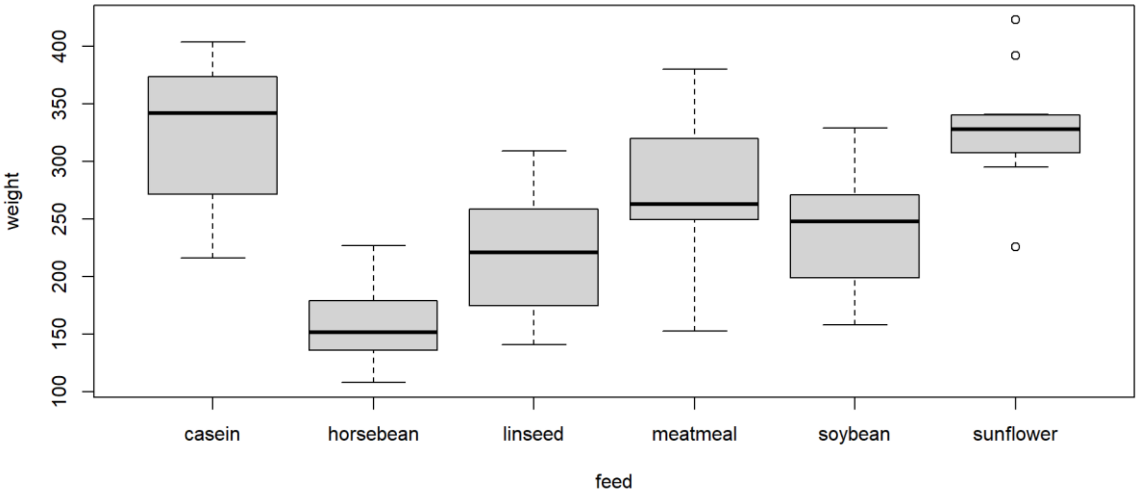
III. ANOVA

A. CHICKWTS

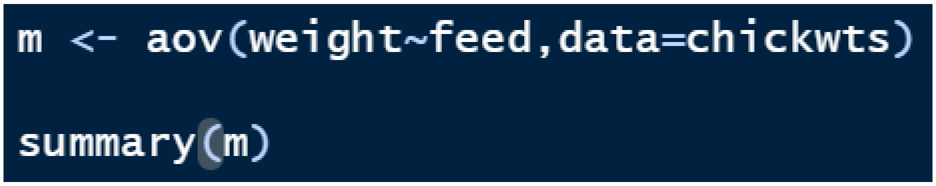
1. BOXPLOT



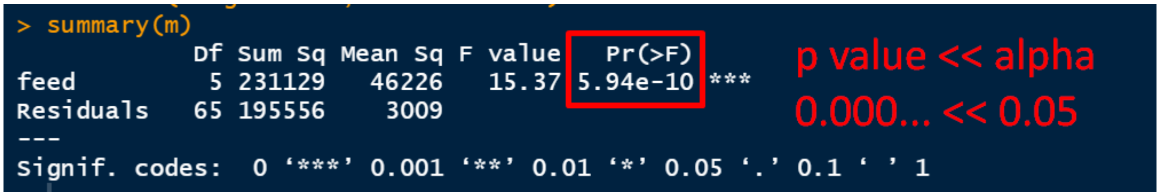
```
boxplot(weight~feed,data=chickwts)
```



2. ANOVA TEST



```
m <- aov(weight~feed,data=chickwts)

summary(m)
```



```
> summary(m)
            Df Sum Sq Mean Sq F value   Pr(>F)    
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p value << alpha

0.000... << 0.05

- H0: The weight of all the chickens are the same → There’s no significant difference between feeding them any of the food.

- H1: The weight of the chickens are significantly different → There's a significant difference between AT LEAST two of the feeds.

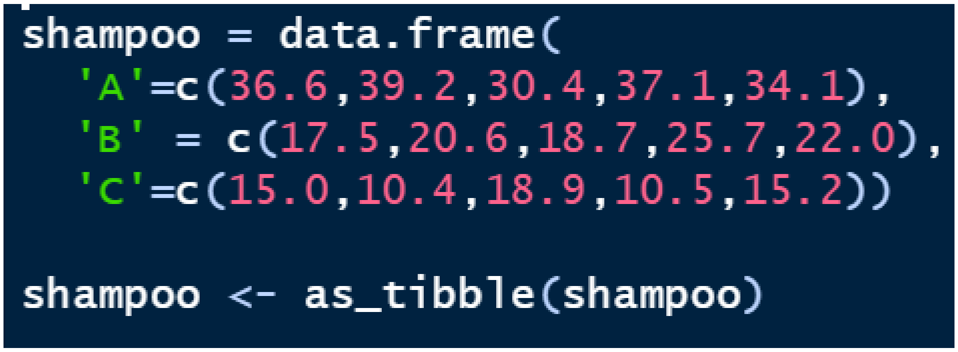
- Since P value << Alpha (0.000… << 0.05) → we accept H1

- Conclusion: there IS a significant difference feeding them the different type of food

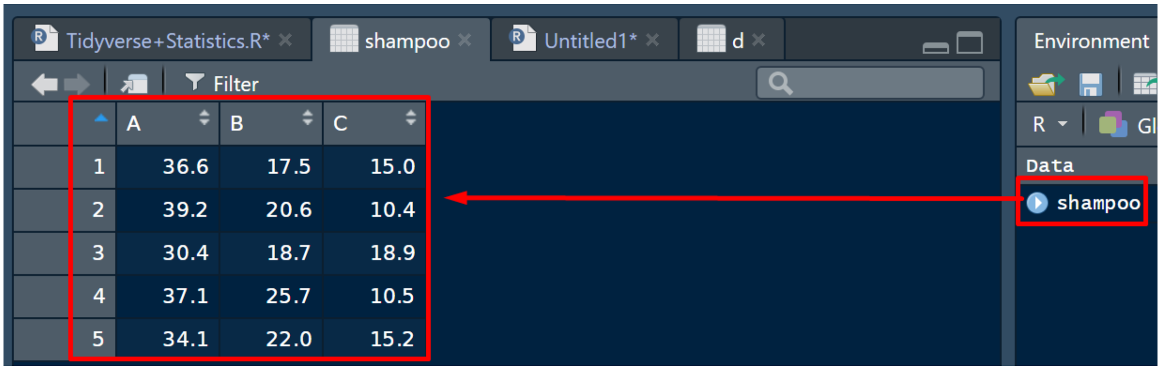
- Most probably is the 'casein' vs ' horsebean' significant difference .

B. SHAMPOO USING %>%

1. CREATING THE DATAFRAME

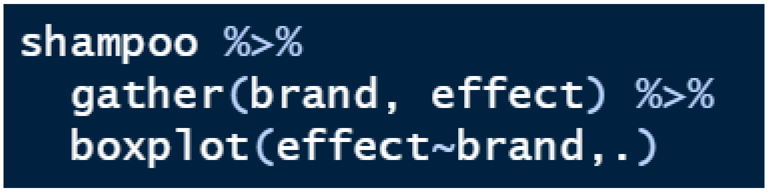


```
shampoo = data.frame(
   'A'=c(36.6,39.2,30.4,37.1,34.1),
   'B' = c(17.5,20.6,18.7,25.7,22.0),
   'C'=c(15.0,10.4,18.9,10.5,15.2))

shampoo <- as_tibble(shampoo)
```

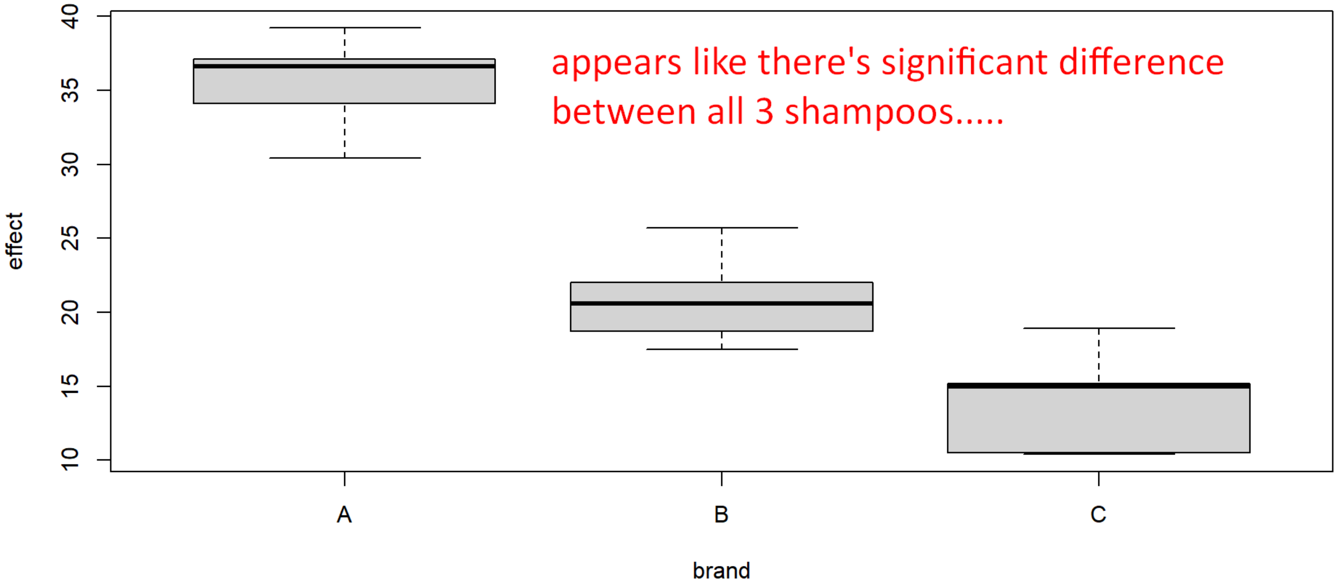


	A	B	C
1	36.6	17.5	15.0
2	39.2	20.6	10.4
3	30.4	18.7	18.9
4	37.1	25.7	10.5
5	34.1	22.0	15.2

2. BOXPLOT

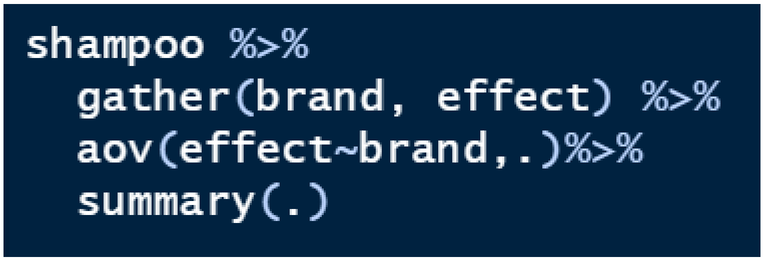


```
shampoo %>%
  gather(brand, effect) %>%
  boxplot(effect~brand,.)
```

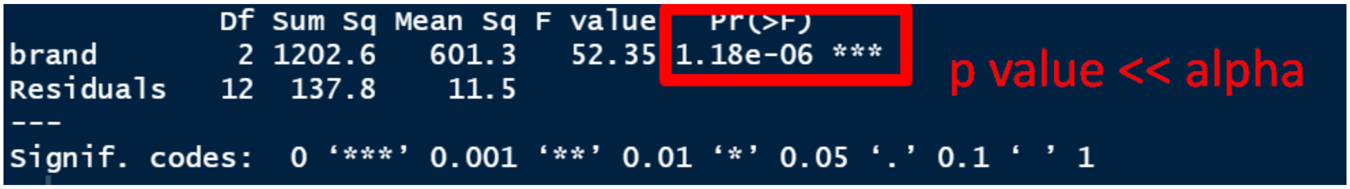


appears like there's significant difference between all 3 shampoos.....

3. ANOVA TEST



```
shampoo %>%
  gather(brand, effect) %>%
  aov(effect~brand,.)%>%
  summary(.)
```



```
            Df Sum Sq Mean Sq F value   Pr(>F)    
brand        2 1202.6   601.3   52.35 1.18e-06 ***
Residuals   12  137.8    11.5                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p value << alpha

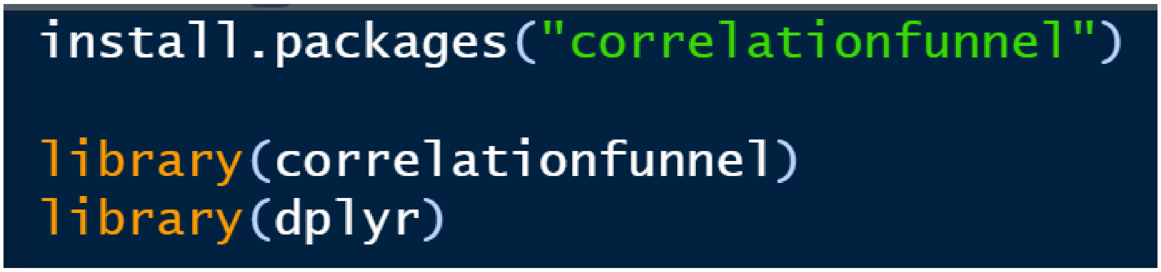
- H0: There’s no significant difference in using all of the shampoos.
- H1: There’s a significant difference in using at least 2 of the shampoos.
- Since P value << Alpha (0.000…. << 0.05) → We accept H1.
- Conclusion: Seems like there’s a difference in effect using Brand A vs B vs C.
- The difference is quite visible from the box plot.

IV. CORRELATION FUNNEL

https://cran.r-project.org/web/packages/correlationfunnel/vignettes/introducing_correlation_funnel.html

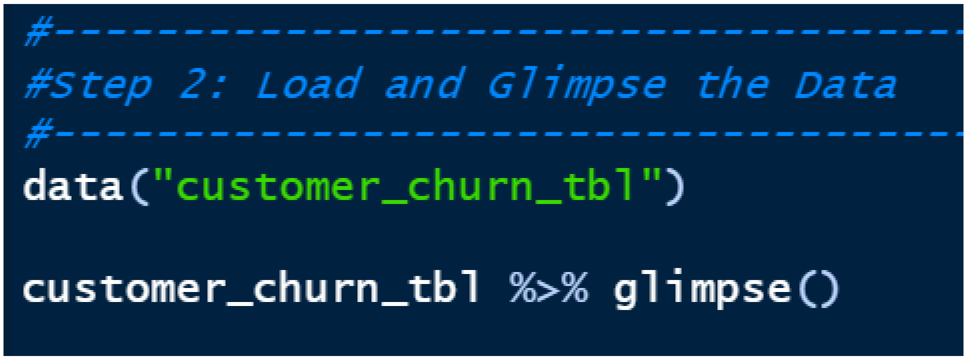
https://www.alvinang.sg/s/correlation-funnel.R

A. STEP 1: INSTALL AND IMPORT PACKAGES

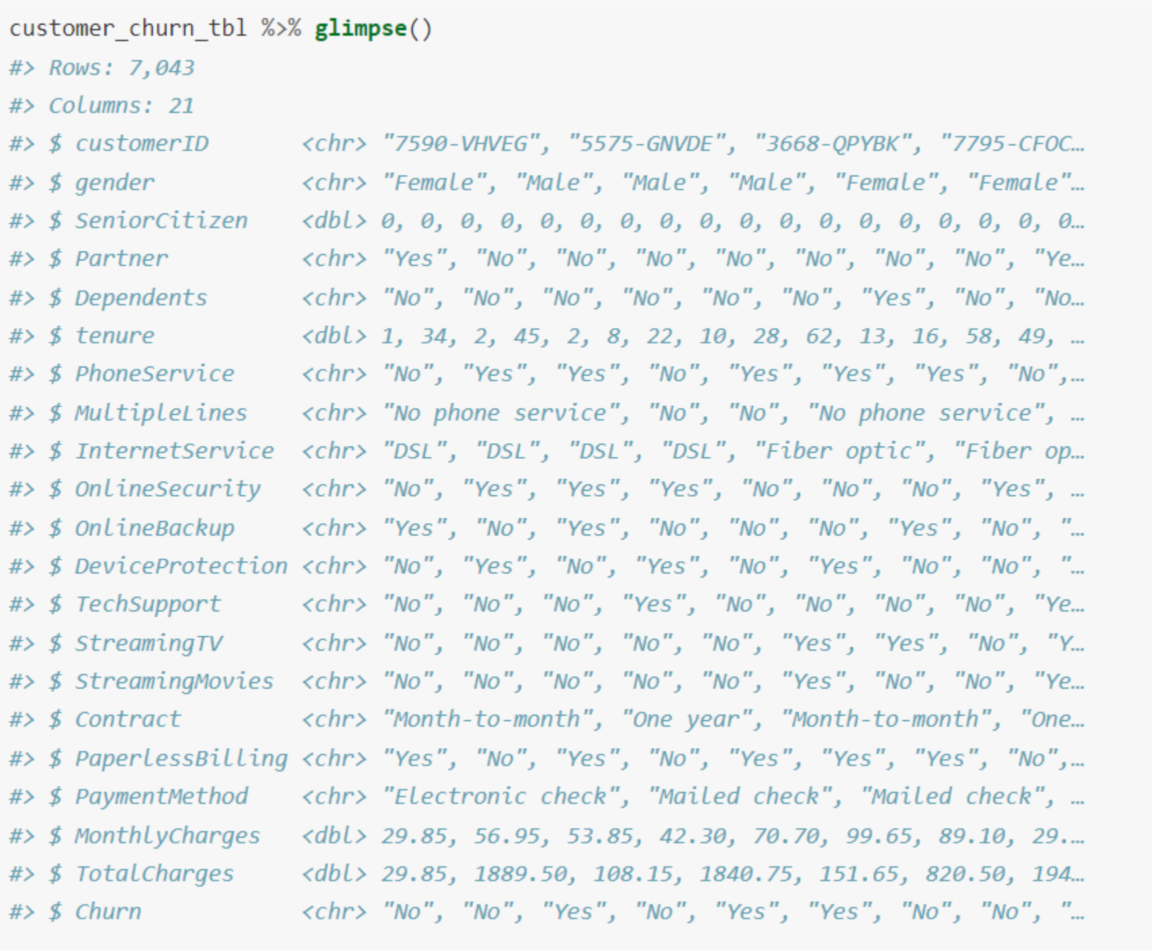


```
install.packages("correlationfunnel")

library(correlationfunnel)
library(dplyr)
```

B. STEP 2: LOAD AND GLIMPSE THE DATA

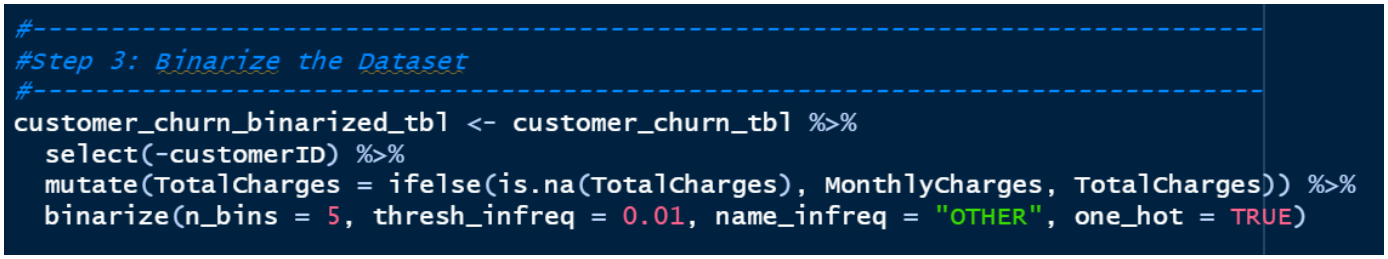


```r
#-----------------------------------
#Step 2: Load and Glimpse the Data
#-----------------------------------
data("customer_churn_tbl")

customer_churn_tbl %>% glimpse()
```



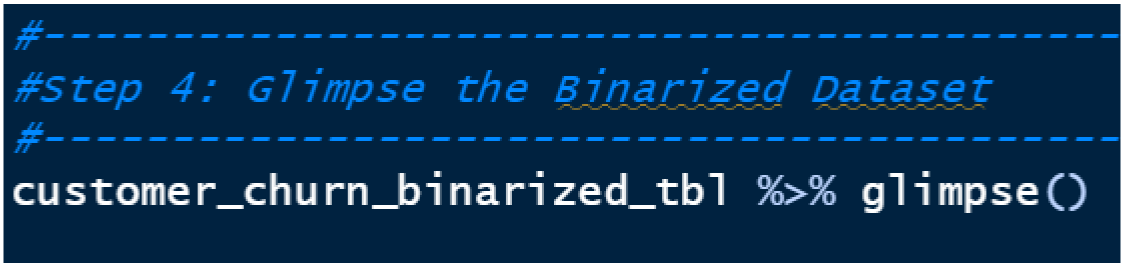
```
customer_churn_tbl %>% glimpse()
#> Rows: 7,043
#> Columns: 21
#> $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOC…
#> $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female"…
#> $ SeniorCitizen    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
#> $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Ye…
#> $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No…
#> $ tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, …
#> $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No",…
#> $ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", …
#> $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber op…
#> $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", …
#> $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "…
#> $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "…
#> $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Ye…
#> $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Y…
#> $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Ye…
#> $ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One…
#> $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No",…
#> $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", …
#> $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.…
#> $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 194…
#> $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "…
```

C. STEP 3: BINARIZE THE DATASET

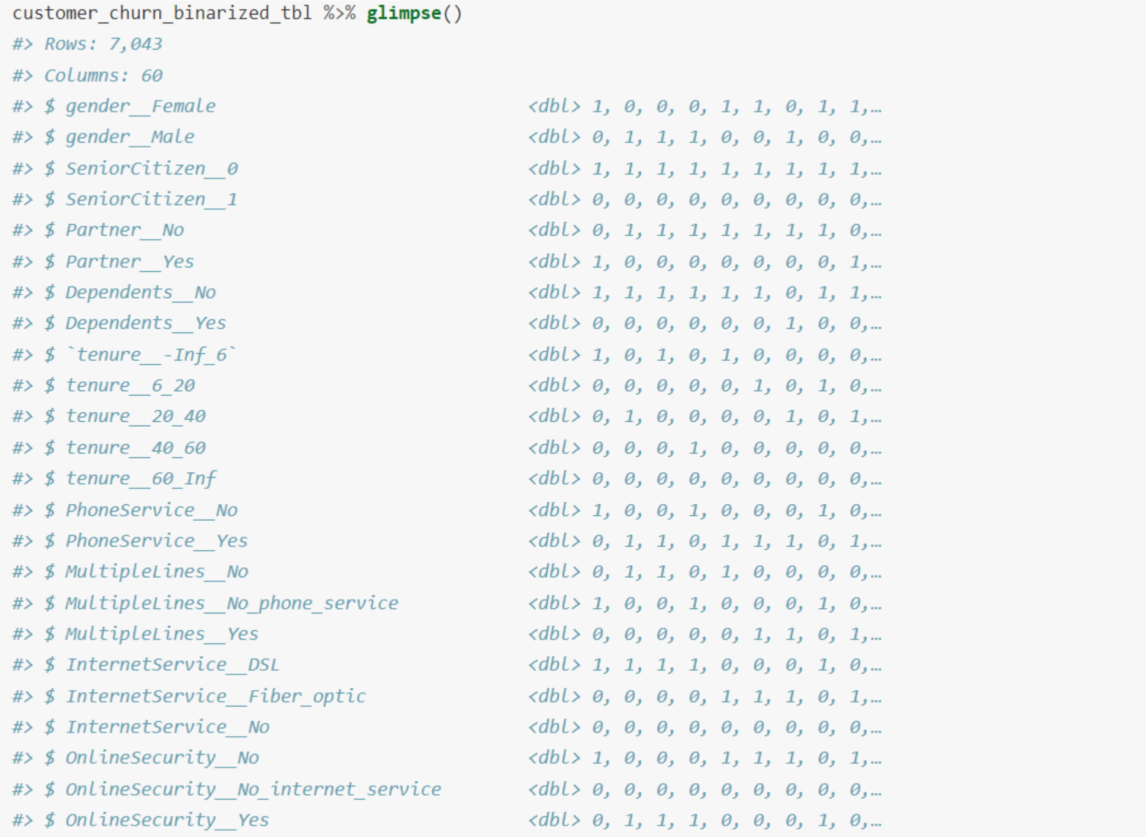


```
#-------------------------------------------------------------------------------
#Step 3: Binarize the Dataset
#-------------------------------------------------------------------------------
customer_churn_binarized_tbl <- customer_churn_tbl %>%
  select(-customerID) %>%
  mutate(TotalCharges = ifelse(is.na(TotalCharges), MonthlyCharges, TotalCharges)) %>%
  binarize(n_bins = 5, thresh_infreq = 0.01, name_infreq = "OTHER", one_hot = TRUE)
```

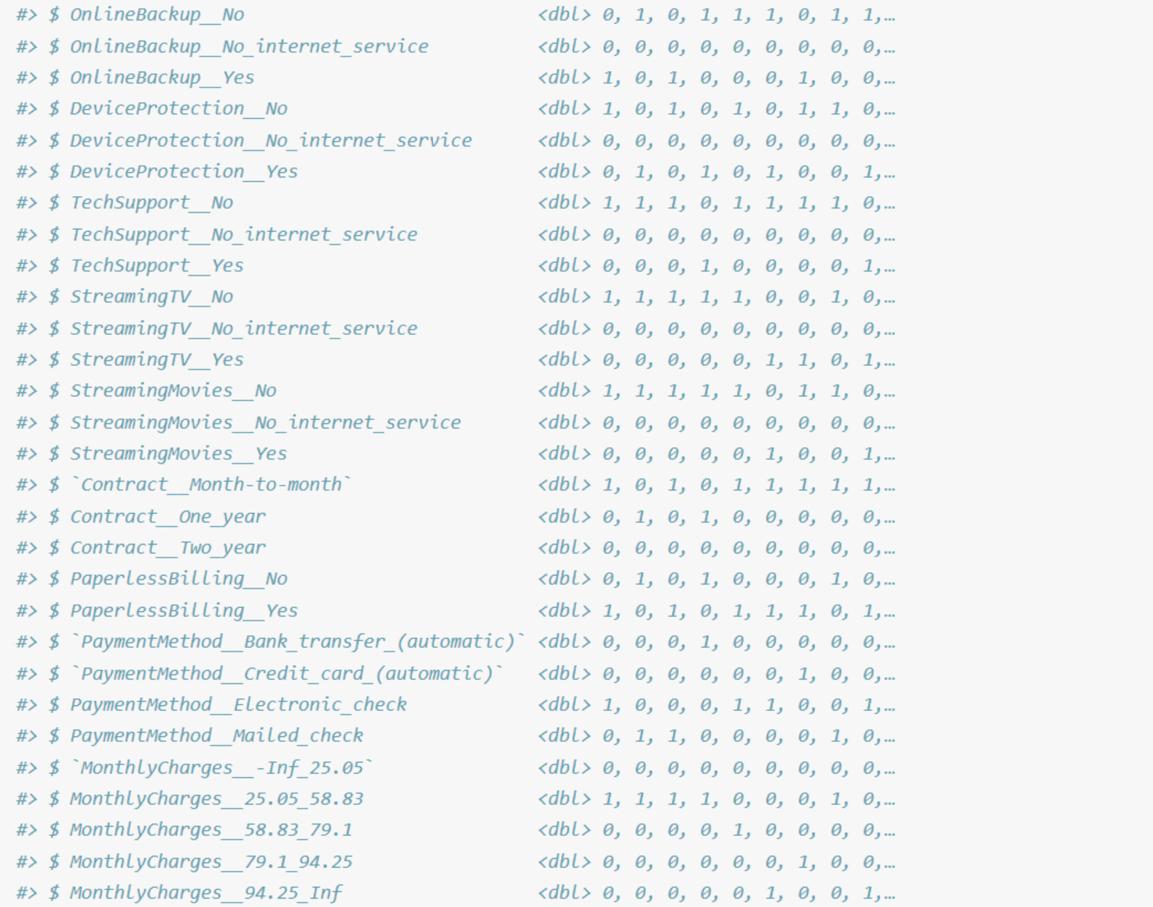
D. STEP 4: GLIMPSE THE BINARIZED DATASET



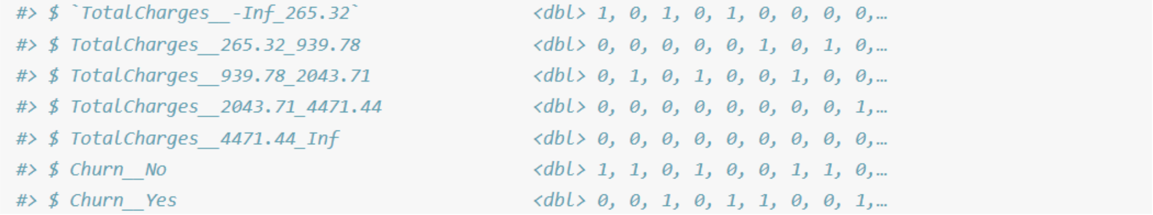
```
#---------------------------------------
#Step 4: Glimpse the Binarized Dataset
#---------------------------------------
customer_churn_binarized_tbl %>% glimpse()
```



```
customer_churn_binarized_tbl %>% glimpse()
#> Rows: 7,043
#> Columns: 60
#> $ gender__Female                         <dbl> 1, 0, 0, 0, 1, 1, 0, 1, 1,…
#> $ gender__Male                           <dbl> 0, 1, 1, 1, 0, 0, 1, 0, 0,…
#> $ SeniorCitizen__0                       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1,…
#> $ SeniorCitizen__1                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ Partner__No                            <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 0,…
#> $ Partner__Yes                           <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1,…
#> $ Dependents__No                         <dbl> 1, 1, 1, 1, 1, 1, 0, 1, 1,…
#> $ Dependents__Yes                        <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0,…
#> $ `tenure__-Inf_6`                       <dbl> 1, 0, 1, 0, 1, 0, 0, 0, 0,…
#> $ tenure__6_20                           <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0,…
#> $ tenure__20_40                          <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 1,…
#> $ tenure__40_60                          <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0,…
#> $ tenure__60_Inf                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ PhoneService__No                       <dbl> 1, 0, 0, 1, 0, 0, 0, 1, 0,…
#> $ PhoneService__Yes                      <dbl> 0, 1, 1, 0, 1, 1, 1, 0, 1,…
#> $ MultipleLines__No                      <dbl> 0, 1, 1, 0, 1, 0, 0, 0, 0,…
#> $ MultipleLines__No_phone_service        <dbl> 1, 0, 0, 1, 0, 0, 0, 1, 0,…
#> $ MultipleLines__Yes                     <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 1,…
#> $ InternetService__DSL                   <dbl> 1, 1, 1, 1, 0, 0, 0, 1, 0,…
#> $ InternetService__Fiber_optic           <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 1,…
#> $ InternetService__No                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ OnlineSecurity__No                     <dbl> 1, 0, 0, 0, 1, 1, 1, 0, 1,…
#> $ OnlineSecurity__No_internet_service    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ OnlineSecurity__Yes                    <dbl> 0, 1, 1, 1, 0, 0, 0, 1, 0,…
```

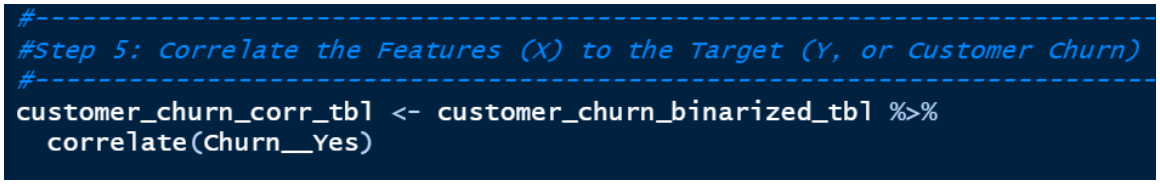


```
#> $ OnlineBackup__No                           <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 1,…
#> $ OnlineBackup__No_internet_service          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ OnlineBackup__Yes                          <dbl> 1, 0, 1, 0, 0, 0, 1, 0, 0,…
#> $ DeviceProtection__No                       <dbl> 1, 0, 1, 0, 1, 0, 1, 1, 0,…
#> $ DeviceProtection__No_internet_service      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ DeviceProtection__Yes                      <dbl> 0, 1, 0, 1, 0, 1, 0, 0, 1,…
#> $ TechSupport__No                            <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 0,…
#> $ TechSupport__No_internet_service           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ TechSupport__Yes                           <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 1,…
#> $ StreamingTV__No                            <dbl> 1, 1, 1, 1, 1, 0, 0, 1, 0,…
#> $ StreamingTV__No_internet_service           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ StreamingTV__Yes                           <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 1,…
#> $ StreamingMovies__No                        <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 0,…
#> $ StreamingMovies__No_internet_service       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ StreamingMovies__Yes                       <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1,…
#> $ `Contract__Month-to-month`                 <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 1,…
#> $ Contract__One_year                         <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 0,…
#> $ Contract__Two_year                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ PaperlessBilling__No                       <dbl> 0, 1, 0, 1, 0, 0, 0, 1, 0,…
#> $ PaperlessBilling__Yes                      <dbl> 1, 0, 1, 0, 1, 1, 1, 0, 1,…
#> $ `PaymentMethod__Bank_transfer_(automatic)` <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0,…
#> $ `PaymentMethod__Credit_card_(automatic)`   <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0,…
#> $ PaymentMethod__Electronic_check            <dbl> 1, 0, 0, 0, 1, 1, 0, 0, 1,…
#> $ PaymentMethod__Mailed_check                <dbl> 0, 1, 1, 0, 0, 0, 0, 1, 0,…
#> $ `MonthlyCharges__-Inf_25.05`               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ MonthlyCharges__25.05_58.83                <dbl> 1, 1, 1, 1, 0, 0, 0, 1, 0,…
#> $ MonthlyCharges__58.83_79.1                 <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0,…
#> $ MonthlyCharges__79.1_94.25                 <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0,…
#> $ MonthlyCharges__94.25_Inf                  <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1,…
```



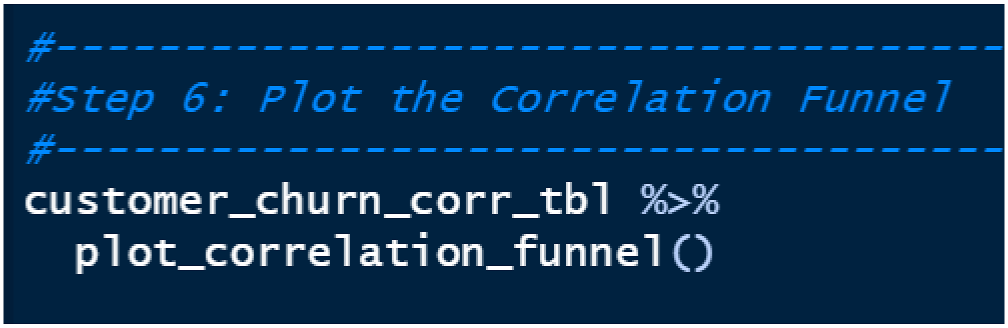
```
#> $ `TotalCharges__-Inf_265.32`               <dbl> 1, 0, 1, 0, 1, 0, 0, 0, 0,…
#> $ TotalCharges__265.32_939.78               <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0,…
#> $ TotalCharges__939.78_2043.71              <dbl> 0, 1, 0, 1, 0, 0, 1, 0, 0,…
#> $ TotalCharges__2043.71_4471.44             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1,…
#> $ TotalCharges__4471.44_Inf                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#> $ Churn__No                                 <dbl> 1, 1, 0, 1, 0, 0, 1, 1, 0,…
#> $ Churn__Yes                                <dbl> 0, 0, 1, 0, 1, 1, 0, 0, 1,…
```

E. STEP 5: CORRELATE THE FEATURES (X) TO THE TARGET (Y, OR CUSTOMER CHURN)

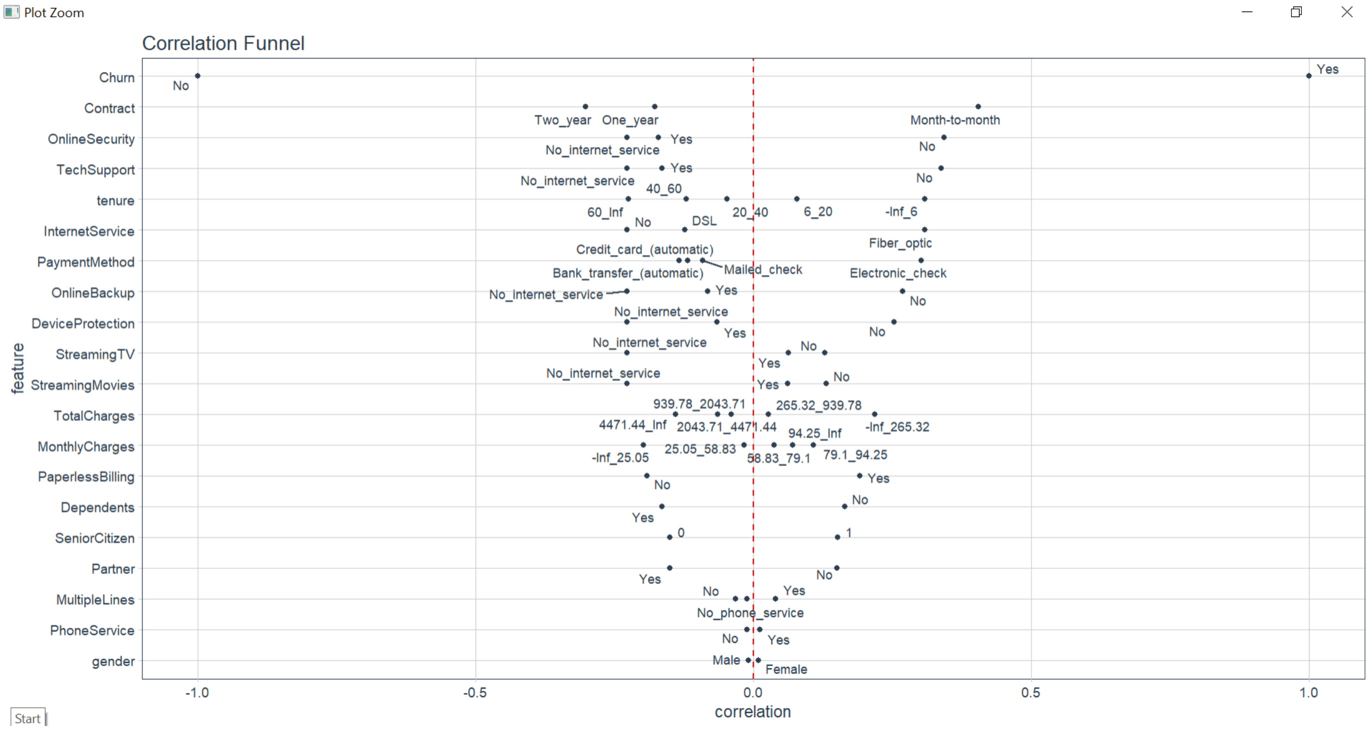


```
#--------------------------------------------------------------------------
#Step 5: Correlate the Features (X) to the Target (Y, or Customer Churn)
#--------------------------------------------------------------------------
customer_churn_corr_tbl <- customer_churn_binarized_tbl %>%
  correlate(Churn__Yes)
```

F. STEP 6: PLOT THE CORRELATION FUNNEL



```
#-----------------------------------
#Step 6: Plot the Correlation Funnel
#-----------------------------------
customer_churn_corr_tbl %>%
  plot_correlation_funnel()
```



G. STEP 7: CONCLUSION (BUSINESS INSIGHTS)

The following features lead to Customers Leaving / Churning:

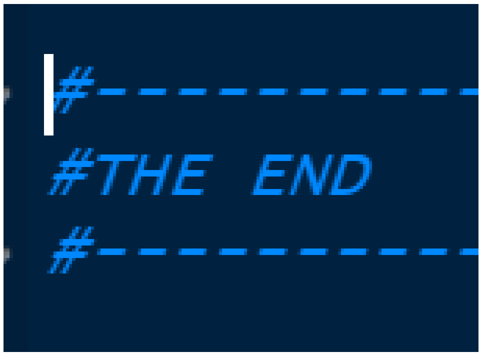
- “Month to Month” Contract Type
- No Online Security
- No Tech Support
- Customer tenure less than 6 months
- Fiber Optic internet service
- Pays with electronic check

The following features lead to Customers Staying (No Churn):

- “Two Year” Contract Type
- Customer Purchases Online Security
- Customer Purchases Tech Support
- Customer tenure greater than 60 months (5 years)
- DSL internet service
- Pays with automatic credit card

We can develop a strategy to retain customers:

- Promotions for 2 Year Contract, Online Security, and Tech Support
- Loyalty Bonuses to incentivize tenure
- Incentives for setting up an automatic credit card payment



```
#-----------
#THE END
#-----------
```

ABOUT THE AUTHOR

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He was a previously a Professor, Scientist and Financial Consultant. Currently, he owns multiple self-started businesses and is a Personal/Business Advisor.

More about him at www.AlvinAng.sg