COLLEGE ADMISSION
Screenshots

project.R × | df ×

Filter

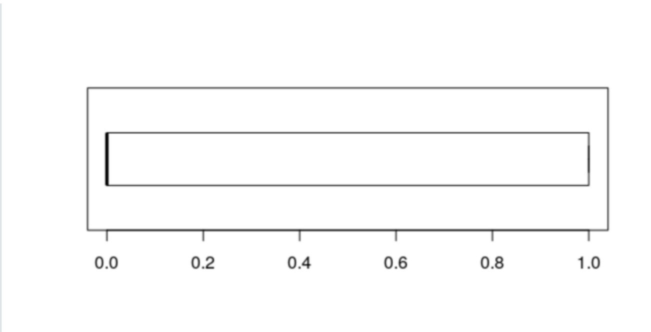
	admit	gre	gpa	ses	Gender_Male	Race	rank
1	0	380	3.61	1	0	3	3
2	1	660	3.67	2	0	2	3
3	1	800	4.00	2	0	2	1
4	1	640	3.19	1	1	2	4
5	0	520	2.93	3	1	2	4
6	1	760	3.00	2	1	1	2
7	1	560	2.98	2	1	2	1
8	0	400	3.08	2	0	2	2

a) To find the null values

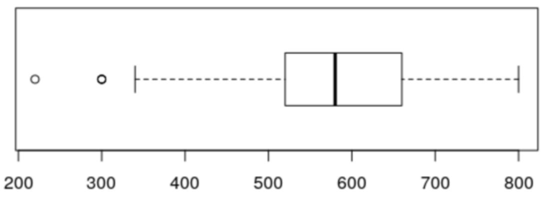
```
> #To find the null values
> sum(is.na(df$admit))
[1] 0
> sum(is.na(df$gre))
[1] 0
> sum(is.na(df$gpa))
[1] 0
> sum(is.na(df$ses))
[1] 0
> sum(is.na(df$Gender_Male))
[1] 0
> sum(is.na(df$Race))
[1] 0
> sum(is.na(df$rank))
[1] 0
>
```

b) To check outliers for

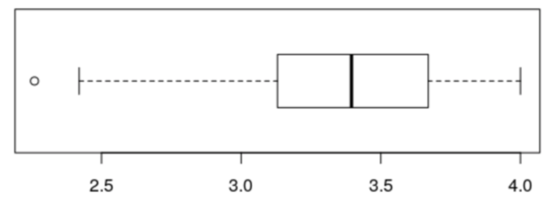
- df$admit



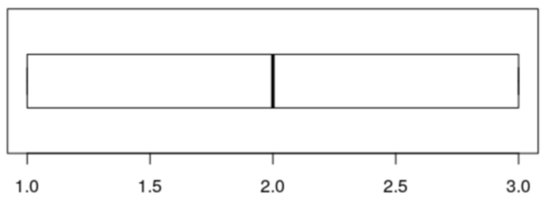
- Df$gre



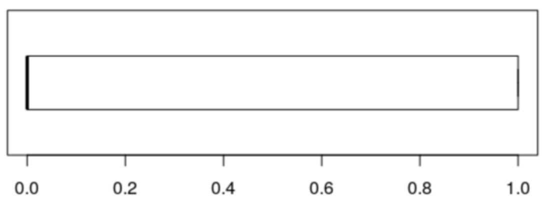
- Df$gpa



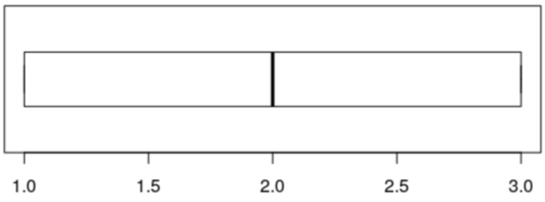
- Df$ses



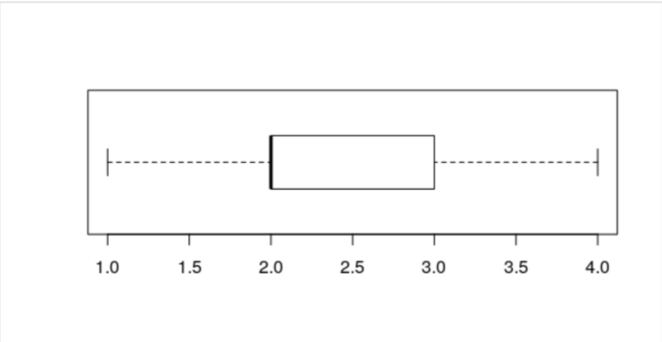
- Df$Gender_Male



- Df$Race

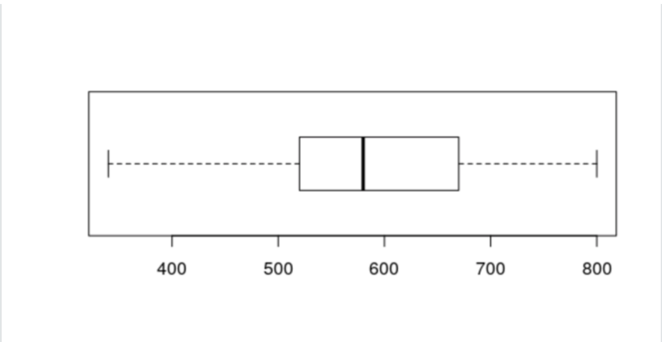


- Df$rank



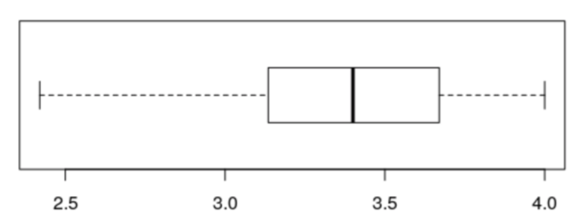
c) Removal Outlier for GRE

```
> #Outlier removal for GRE
> outliers1 <- boxplot(df$gre, plot=FALSE)$out
> df[which(df$gre %in% outliers1),]
    admit gre  gpa ses Gender_Male Race rank
72      0 300 2.92   1           1    1    4
180     0 300 3.01   2           0    1    3
305     0 220 2.83   1           1    3    3
316     1 300 2.84   3           1    1    2
> df <- df[-which(df$gre %in% outliers1),]
> boxplot(df$gre, horizontal = TRUE)
```



d) Removal of Outlier for GPA

```
> #Outlier removed for GPA
> outliers2 <- boxplot(df$gpa, plot=FALSE)$out
> df[which(df$gpa %in% outliers2),]
    admit gre  gpa ses Gender_Male Race rank
290     0 420 2.26   2           1    2    4
> df <- df[-which(df$gpa %in% outliers2),]
> boxplot(df$gpa, horizontal = TRUE)
>
```



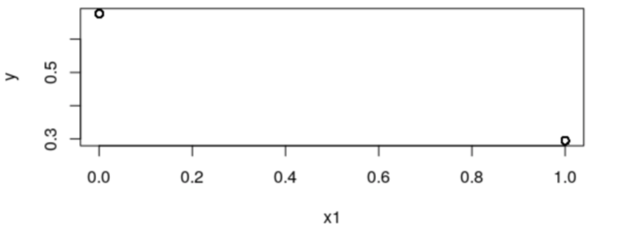
e) Conversion of data type

```
> #to find the structure of the dataset
> str(df)
'data.frame':	395 obs. of  7 variables:
 $ admit      : int  0 1 1 1 0 1 1 0 1 0 ...
 $ gre        : int  380 660 800 640 520 760 560 400 540 700 ...
 $ gpa        : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
 $ ses        : int  1 2 2 1 3 2 2 2 1 1 ...
 $ Gender_Male: int  0 0 0 1 1 1 1 0 1 0 ...
 $ Race       : int  3 2 2 2 2 1 2 2 1 2 ...
 $ rank       : int  3 3 1 4 4 2 1 2 3 2 ...
```

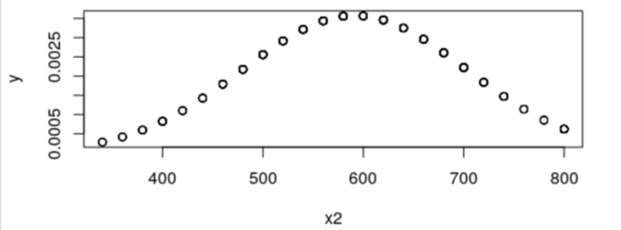
```
> str(df)
'data.frame':	395 obs. of  9 variables:
 $ admit      : int  0 1 1 1 0 1 1 0 1 0 ...
 $ gre        : int  380 660 800 640 520 760 560 400 540 700 ...
 $ gpa        : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
 $ ses        : Factor w/ 3 levels "1","2","3": 1 2 2 1 3 2 2 2 1 1 ...
 $ Gender_Male: Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 1 2 1 ...
 $ Race       : Factor w/ 3 levels "1","2","3": 3 2 2 2 2 1 2 2 1 2 ...
 $ rank       : Factor w/ 4 levels "1","2","3","4": 3 3 1 4 4 2 1 2 3 2 ...
 $ gpa_fac    : Factor w/ 129 levels "2.42","2.48",..: 93 99 129 52 27 33 32 41 71 122 ...
 $ admit_fac  : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 1 2 1 ...
>
```

f) Normal Distribution

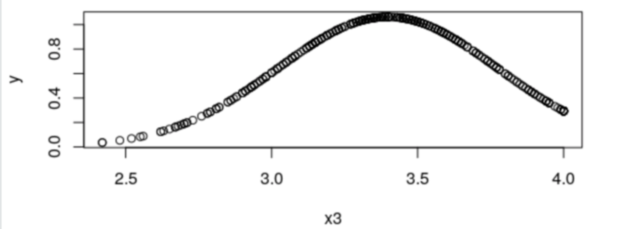
- Admit



- GRE



- GPA



g) Correlation

```
> round(res$P, 3)
            admit   gre   gpa   ses Gender_Male  Race  rank gpa_fac admit_fac
admit          NA 0.000 0.000 0.265       0.615 0.246 0.000   0.000     0.000
gre         0.000    NA 0.000 0.374       0.583 0.221 0.032   0.000     0.000
gpa         0.000 0.000    NA 0.950       0.990 0.286 0.429   0.000     0.000
ses         0.265 0.374 0.950    NA       0.614 0.323 0.744   0.950     0.265
Gender_Male 0.615 0.583 0.990 0.614          NA 0.277 0.486   0.990     0.615
Race        0.246 0.221 0.286 0.323       0.277    NA 0.429   0.286     0.246
rank        0.000 0.032 0.429 0.744       0.486 0.429    NA   0.429     0.000
gpa_fac     0.000 0.000 0.000 0.950       0.990 0.286 0.429      NA     0.000
admit_fac   0.000 0.000 0.000 0.265       0.615 0.246 0.000   0.000        NA
```

Df_2

	gre	gpa	rank	admit_fac
1	380	3.61	3	0
2	660	3.67	3	1
3	800	4.00	1	1
4	640	3.19	4	1
5	520	2.93	4	0
6	760	3.00	2	1
7	560	2.98	1	1
8	400	3.08	2	0
9	540	3.39	3	1
10	700	3.92	2	0
11	800	4.00	4	0

h) Logistic Regression

```
Call:
glm(formula = admit_fac ~ ., family = binomial(link = "logit"),
    data = train, model = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6510  -0.8584  -0.5997   1.0179   2.1923

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.078683   1.433223  -2.846 0.004430 **
gre          0.003118   0.001388   2.247 0.024664 *
gpa          0.772063   0.408796   1.889 0.058942 .
rank2       -1.024282   0.406347  -2.521 0.011712 *
rank3       -1.700077   0.451925  -3.762 0.000169 ***
rank4       -1.970466   0.529978  -3.718 0.000201 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 345.55  on 275  degrees of freedom
Residual deviance: 308.89  on 270  degrees of freedom
AIC: 320.89

Number of Fisher Scoring iterations: 4
```

	gre	gpa	rank	admit_fac	admit_prob	admit_pred
2	660	3.67	3	1	0.29160663	0
4	640	3.19	4	1	0.16925332	0
5	520	2.93	4	0	0.10286481	0
7	560	2.98	1	1	0.49199753	1
8	400	3.08	2	0	0.18574130	0
13	760	4.00	1	1	0.79883603	1
22	660	3.63	2	1	0.43962552	1
23	600	2.82	4	0	0.11906645	0
31	540	3.78	4	0	0.19043551	0
32	760	3.35	3	0	0.30515218	0
33	600	3.40	3	0	0.21702063	0

i) Confusion Matrix

```
Confusion Matrix and Statistics


    0  1
 0 62 19
 1 20 18

               Accuracy : 0.6723
                 95% CI : (0.5802, 0.7555)
    No Information Rate : 0.6891
    P-Value [Acc > NIR] : 0.693

                  Kappa : 0.2408

 Mcnemar's Test P-Value : 1.000

            Sensitivity : 0.7561
            Specificity : 0.4865
         Pos Pred Value : 0.7654
         Neg Pred Value : 0.4737
             Prevalence : 0.6891
         Detection Rate : 0.5210
   Detection Prevalence : 0.6807
      Balanced Accuracy : 0.6213

       'Positive' Class : 0
```

j) SVM Model

```
> summary(svm_model1)

Call:
svm(formula = admit_fac ~ ., data = train1)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  188

 ( 86 102 )


Number of Classes:  2

Levels:
 0 1
```

project.R* test1 df_3 test df_2 df

Filter

	gre	gpa	rank	admit_fac	prediction_svm
2	660	3.67	3	1	0
4	640	3.19	4	1	0
5	520	2.93	4	0	0
7	560	2.98	1	1	0
8	400	3.08	2	0	0
13	760	4.00	1	1	1
22	660	3.63	2	1	0
23	600	2.82	4	0	0
31	540	3.78	4	0	0
32	760	3.35	3	0	0
33	600	3.40	3	0	0

k) Confusion Matrix

```
Confusion Matrix and Statistics


     0  1
  0 73  8
  1 30  8

               Accuracy : 0.6807
                 95% CI : (0.589, 0.7631)
    No Information Rate : 0.8655
    P-Value [Acc > NIR] : 0.9999999

                  Kappa : 0.1321

 Mcnemar's Test P-Value : 0.0006577

            Sensitivity : 0.7087
            Specificity : 0.5000
         Pos Pred Value : 0.9012
         Neg Pred Value : 0.2105
             Prevalence : 0.8655
         Detection Rate : 0.6134
   Detection Prevalence : 0.6807
      Balanced Accuracy : 0.6044

       'Positive' Class : 0
```

l) Decision Tree

```
> summary(dt_model)
Call:
rpart(formula = admit_fac ~ ., data = train2, method = "class")
  n= 276

          CP nsplit rel error    xerror       xstd
1 0.09090909      0 1.0000000 1.0000000 0.08797982
2 0.01893939      2 0.8181818 0.8750000 0.08467100
3 0.01000000      8 0.6704545 0.8863636 0.08500465

Variable importance
 gpa rank  gre
  56   26   18

Node number 1: 276 observations,    complexity param=0.09090909
  predicted class=0  expected loss=0.3188406  P(node) =1
    class counts:   188    88
   probabilities: 0.681 0.319
  left son=2 (240 obs) right son=3 (36 obs)
  Primary splits:
      rank splits as  RLLL,      improve=7.072947, (0 missing)
      gpa  < 3.435 to the left,  improve=7.057937, (0 missing)
      gre  < 510   to the left,  improve=5.089228, (0 missing)
```

```
Node number 2: 240 observations,    complexity param=0.01893939
  predicted class=0  expected loss=0.275  P(node) =0.8695652
    class counts:   174    66
   probabilities: 0.725 0.275
  left son=4 (60 obs) right son=5 (180 obs)
  Primary splits:
      gre  < 510   to the left,  improve=4.011111, (0 missing)
      rank splits as  -RLL,      improve=3.111863, (0 missing)
      gpa  < 3.435 to the left,  improve=2.664253, (0 missing)
  Surrogate splits:
      gpa < 2.675 to the left,  agree=0.758, adj=0.033, (0 split)

Node number 3: 36 observations,    complexity param=0.09090909
  predicted class=1  expected loss=0.3888889  P(node) =0.1304348
    class counts:    14    22
   probabilities: 0.389 0.611
  left son=6 (16 obs) right son=7 (20 obs)
  Primary splits:
      gpa < 3.4   to the left,  improve=7.511111, (0 missing)
      gre < 670   to the left,  improve=1.017466, (0 missing)
  Surrogate splits:
      gre < 450   to the left,  agree=0.667, adj=0.25, (0 split)
```

```
Node number 4: 60 observations
  predicted class=0  expected loss=0.1166667  P(node) =0.2173913
    class counts:    53     7
   probabilities: 0.883 0.117

Node number 5: 180 observations,    complexity param=0.01893939
  predicted class=0  expected loss=0.3277778  P(node) =0.6521739
    class counts:   121    59
   probabilities: 0.672 0.328
  left son=10 (87 obs) right son=11 (93 obs)
  Primary splits:
      rank splits as  -RLL,      improve=1.889519, (0 missing)
      gpa  < 3.37  to the left,  improve=1.425079, (0 missing)
      gre  < 650   to the left,  improve=1.039757, (0 missing)
  Surrogate splits:
      gre < 570   to the left,  agree=0.567, adj=0.103, (0 split)
      gpa < 3.275 to the right, agree=0.550, adj=0.069, (0 split)

Node number 6: 16 observations
  predicted class=0  expected loss=0.25  P(node) =0.05797101
    class counts:    12     4
   probabilities: 0.750 0.250
```

```
Node number 7: 20 observations
  predicted class=1  expected loss=0.1  P(node) =0.07246377
    class counts:     2    18
   probabilities: 0.100 0.900

Node number 10: 87 observations
  predicted class=0  expected loss=0.2528736  P(node) =0.3152174
    class counts:    65    22
   probabilities: 0.747 0.253

Node number 11: 93 observations,    complexity param=0.01893939
  predicted class=0  expected loss=0.3978495  P(node) =0.3369565
    class counts:    56    37
   probabilities: 0.602 0.398
  left son=22 (86 obs) right son=23 (7 obs)
  Primary splits:
      gpa < 3.945 to the left,  improve=3.193691, (0 missing)
      gre < 710   to the left,  improve=1.991239, (0 missing)
```

```
Node number 22: 86 observations,    complexity param=0.01893939
  predicted class=0  expected loss=0.3604651  P(node) =0.3115942
    class counts:    55    31
   probabilities: 0.640 0.360
  left son=44 (8 obs) right son=45 (78 obs)
  Primary splits:
      gpa < 2.935 to the left,  improve=2.292188, (0 missing)
      gre < 710   to the left,  improve=1.298531, (0 missing)

Node number 23: 7 observations
  predicted class=1  expected loss=0.1428571  P(node) =0.02536232
    class counts:     1     6
   probabilities: 0.143 0.857

Node number 44: 8 observations
  predicted class=0  expected loss=0  P(node) =0.02898551
    class counts:     8     0
   probabilities: 1.000 0.000
```

```
Node number 45: 78 observations,    complexity param=0.01893939
  predicted class=0  expected loss=0.3974359  P(node) =0.2826087
    class counts:    47    31
   probabilities: 0.603 0.397
  left son=90 (33 obs) right son=91 (45 obs)
  Primary splits:
      gpa < 3.495 to the right, improve=1.779176, (0 missing)
      gre < 710   to the left,  improve=1.474916, (0 missing)
  Surrogate splits:
      gre < 690   to the right, agree=0.628, adj=0.121, (0 split)

Node number 90: 33 observations
  predicted class=0  expected loss=0.2727273  P(node) =0.1195652
    class counts:    24     9
   probabilities: 0.727 0.273

Node number 91: 45 observations,    complexity param=0.01893939
  predicted class=0  expected loss=0.4888889  P(node) =0.1630435
    class counts:    23    22
   probabilities: 0.511 0.489
  left son=182 (35 obs) right son=183 (10 obs)
  Primary splits:
      gpa < 3.41  to the left,  improve=4.3460320, (0 missing)
      gre < 650   to the left,  improve=0.9199234, (0 missing)
```

```
Node number 182: 35 observations
  predicted class=0  expected loss=0.3714286  P(node) =0.1268116
    class counts:    22    13
   probabilities: 0.629 0.371

Node number 183: 10 observations
  predicted class=1  expected loss=0.1  P(node) =0.03623188
    class counts:     1     9
   probabilities: 0.100 0.900
```

m) Confusion Matrix

```
Confusion Matrix and Statistics


     0  1
  0 21 60
  1  6 32

               Accuracy : 0.4454
                 95% CI : (0.3543, 0.5393)
    No Information Rate : 0.7731
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0736

 Mcnemar's Test P-Value : 6.853e-11

            Sensitivity : 0.7778
            Specificity : 0.3478
         Pos Pred Value : 0.2593
         Neg Pred Value : 0.8421
             Prevalence : 0.2269
         Detection Rate : 0.1765
   Detection Prevalence : 0.6807
      Balanced Accuracy : 0.5628

       'Positive' Class : 0
```