Column Categorization and Principal Component Analysis(PCA)

```
> names(loan)
  [1] "X"                          "loan_amnt"
  [3] "funded_amnt"                "funded_amnt_inv"
  [5] "term"                       "int_rate"
  [7] "installment"                "grade"
  [9] "sub_grade"                  "emp_title"
 [11] "emp_length"                 "home_ownership"
 [13] "annual_inc"                 "verification_status"
 [15] "issue_d"                    "loan_status"
 [17] "pymnt_plan"                 "desc"
 [19] "purpose"                    "title"
 [21] "zip_code"                   "addr_state"
 [23] "dti"                        "delinq_2yrs"
 [25] "earliest_cr_line"           "inq_last_6mths"
 [27] "mths_since_last_delinq"     "open_acc"
 [29] "pub_rec"                    "revol_bal"
 [31] "revol_util"                 "total_acc"
 [33] "initial_list_status"        "out_prncp"
 [35] "out_prncp_inv"              "total_pymnt"
 [37] "total_pymnt_inv"            "total_rec_prncp"
 [39] "total_rec_int"              "total_rec_late_fee"
 [41] "recoveries"                 "collection_recovery_fee"
 [43] "last_pymnt_d"               "last_pymnt_amnt"
 [45] "next_pymnt_d"               "last_credit_pull_d"
 [47] "collections_12_mths_ex_med" "policy_code"
 [49] "application_type"           "verification_status_joint"
 [51] "acc_now_delinq"             "tot_coll_amt"
 [53] "tot_cur_bal"                "open_acc_6m"
 [55] "open_act_il"                "open_il_12m"
 [57] "open_il_24m"                "mths_since_rcnt_il"
 [59] "total_bal_il"               "il_util"
 [61] "open_rv_12m"                "open_rv_24m"
 [63] "max_bal_bc"                 "all_util"
 [65] "total_rev_hi_lim"           "inq_fi"
 [67] "total_cu_tl"                "inq_last_12m"
 [69] "acc_open_past_24mths"       "avg_cur_bal"
 [71] "bc_open_to_buy"             "bc_util"
 [73] "chargeoff_within_12_mths"   "delinq_amnt"
 [75] "mo_sin_old_il_acct"         "mo_sin_old_rev_tl_op"
 [77] "mo_sin_rcnt_rev_tl_op"      "mo_sin_rcnt_tl"
 [79] "mort_acc"                   "mths_since_recent_bc"
 [81] "mths_since_recent_inq"      "num_accts_ever_120_pd"
 [83] "num_actv_bc_tl"             "num_actv_rev_tl"
 [85] "num_bc_sats"                "num_bc_tl"
 [87] "num_il_tl"                  "num_op_rev_tl"
 [89] "num_rev_accts"              "num_rev_tl_bal_gt_0"
 [91] "num_sats"                   "num_tl_120dpd_2m"
 [93] "num_tl_30dpd"               "num_tl_90g_dpd_24m"
 [95] "num_tl_op_past_12m"         "pct_tl_nvr_dlq"
 [97] "percent_bc_gt_75"           "pub_rec_bankruptcies"
 [99] "tax_liens"                  "tot_hi_cred_lim"
[101] "total_bal_ex_mort"          "total_bc_limit"
[103] "total_il_high_credit_limit" "sec_app_earliest_cr_line"
[105] "hardship_flag"              "hardship_type"
[107] "hardship_reason"            "hardship_status"
[109] "hardship_start_date"        "hardship_end_date"
[111] "payment_plan_start_date"    "hardship_loan_status"
[113] "disbursement_method"        "debt_settlement_flag"
[115] "debt_settlement_flag_date"  "settlement_status"
[117] "settlement_date"
```

```
> View(loan)
```

loan

Filter Cols: « ‹ 1 - 50 › »

	X	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	annual_inc	verification_st
1	186	4500	4500	4500	36 months	11.31	147.99	B	B3	Accounts Examiner III	10+ years	RENT	38500.00	Not Verifie
2	296	2500	2500	2475	36 months	13.56	84.92	C	C1	Manager	5 years	RENT	42000.00	Not Verifie
3	369	4000	4000	4000	36 months	17.97	144.55	D	D1	service advisor	5 years	MORTGAGE	60000.00	Source Ver
4	402	1000	1000	1000	36 months	23.40	38.92	E	E1	HR Director	3 years	RENT	60000.00	Source Ver
5	510	5000	5000	5000	36 months	7.56	155.67	A	A3	Supply chain management	5 years	MORTGAGE	98000.00	Not Verifie
6	800	10000	10000	10000	60 months	12.98	227.43	B	B5	kitchen and bath designer	1 year	MORTGAGE	60000.00	Not Verifie
7	829	29050	29050	29050	36 months	10.33	941.87	B	B1	Executive Casino Host	< 1 year	MORTGAGE	68000.00	Source Ver
8	835	1000	1000	1000	36 months	13.56	33.97	C	C1	Customer service	1 year	RENT	42140.00	Source Ver
9	930	10000	10000	10000	36 months	11.80	331.19	B	B4	Director of Maintenance	1 year	RENT	100000.00	Source Ver
10	1066	10000	10000	10000	36 months	8.19	314.25	A	A4	Teacher and Coach	8 years	MORTGAGE	65000.00	Verified
11	1104	9500	9500	9500	36 months	8.19	298.53	A	A4	Table Games Dealer	5 years	RENT	50000.00	Not Verifie
12	1527	10000	10000	10000	36 months	14.47	344.07	C	C2	operator	10+ years	RENT	80000.00	Not Verifie
13	1875	10000	10000	10000	36 months	8.19	314.25	A	A4	Customer Service Agent	10+ years	OWN	50000.00	Source Ver
14	1953	5000	5000	5000	36 months	10.33	162.12	B	B1	Police Officer	10+ years	MORTGAGE	118964.00	Not Verifie
15	2175	3000	3000	3000	36 months	19.92	111.37	D	D3	Teacher	1 year	RENT	62000.00	Not Verifie
16	2787	1000	1000	1000	36 months	27.27	40.98	E	E5	Warehouse Associate	5 years	RENT	40000.00	Source Ver
17	3002	11000	11000	11000	36 months	10.72	358.67	B	B2	Sr. Mechanic	10+ years	MORTGAGE	125000.00	Source Ver
18	3177	11000	11000	10750	36 months	12.98	370.53	B	B5	Agent	10+ years	MORTGAGE	130000.00	Source Ver
19	3765	6500	6500	6500	36 months	8.81	206.13	A	A5	OFFICE MANAGER AND LEASE ADMINISTRATOR	3 years	MORTGAGE	43000.00	Source Ver
20	3992	13800	13800	13800	60 months	26.31	415.72	E	E4	Investigative Specialist	8 years	MORTGAGE	71000.00	Not Verifie
21	4384	14950	14950	14950	36 months	23.40	581.84	E	E1	Branch manager	3 years	MORTGAGE	70000.00	Source Ver
22	4414	8000	8000	8000	36 months	12.98	269.48	B	B5	Carpenter	10+ years	OWN	89675.00	Not Verifie

```
> loan = loan %>% select(loan_status , loan_amnt , funded_amnt, installment,
int_rate, issue_d , grade , purpose, dti,
+            emp_length , home_ownership ,annual_inc , term)
```

mvap1.Rmd × datacleaning.r × New_mvap.R* × loan × Untitled2* × MVA DAY 1 code.R ×

Filter

	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	purpose	dti
1	Fully Paid	4500	4500	147.99	11.31	Dec-2018	B	credit_card	
2	Fully Paid	2500	2500	84.92	13.56	Dec-2018	C	other	
3	Fully Paid	4000	4000	144.55	17.97	Dec-2018	D	house	
4	Fully Paid	1000	1000	38.92	23.40	Dec-2018	E	debt_consolidation	
5	Fully Paid	5000	5000	155.67	7.56	Dec-2018	A	credit_card	
6	Fully Paid	10000	10000	227.43	12.98	Dec-2018	B	car	
7	Fully Paid	29050	29050	941.87	10.33	Dec-2018	B	home_improvement	
8	Fully Paid	1000	1000	33.97	13.56	Dec-2018	C	moving	
9	Fully Paid	10000	10000	331.19	11.80	Dec-2018	B	debt_consolidation	
10	Fully Paid	10000	10000	314.25	8.19	Dec-2018	A	debt_consolidation	
11	Fully Paid	9500	9500	298.53	8.19	Dec-2018	A	credit_card	
12	Fully Paid	10000	10000	344.07	14.47	Dec-2018	C	debt_consolidation	

Binarization of Term column (36 <- 1 and 60 <- 0)

```
> unique(loan$term)
[1]  36 months  60 months
Levels:  36 months  60 months

> loan$term <- as.integer(gsub("months", "", loan$term))

> loan$term[loan$term == 36] <-  1

> loan$term[loan$term != 1] <- 0

> unique(loan$term)
[1] 1 0
```

Categorization of grade

```
> unique(loan$grade)
[1] B C D E A G F
Levels: A B C D E F G

> loan$grade <- as.character(loan$grade)

> loan$grade[loan$grade == "A"] <- 7

> loan$grade[loan$grade == "B"] <- 6

> loan$grade[loan$grade == "C"] <- 5

> loan$grade[loan$grade == "D"] <- 4

> loan$grade[loan$grade == "E"] <- 3

> loan$grade[loan$grade == "F"] <- 2

> loan$grade[loan$grade == "G"] <- 1

> loan$grade <- as.integer(loan$grade)

> unique(loan$grade)
[1] 6 5 4 3 7 1 2
```

Clearance of emp_length variable

```
> unique(loan$emp_length)
 [1] "10+ years" "5 years"   "3 years"   "1 year"    " 1 year"   "8 years"
 [7] "2 years"   "7 years"   "4 years"   ""          "6 years"   "9 years"

> loan$emp_length <- gsub("<", "", loan$emp_length)

> loan$emp_length <- gsub("years", "", loan$emp_length)

> loan$emp_length <- gsub("year", "", loan$emp_length)

> loan$emp_length <- gsub("n/a", "", loan$emp_length)

> loan$emp_length <- gsub(" ", "", loan$emp_length)

> loan$emp_length <- gsub("\\+", "", loan$emp_length)
```

```
> loan$emp_length <- ifelse(loan$emp_length =="", 10, loan$emp_length)

> loan$emp_length <- as.integer(loan$emp_length)

> unique(loan$emp_length)
 [1] 10  5  3  1  8  2  7  4  6  9
```

Binarization of home_ownership

```
> unique(loan$home_ownership)
[1] RENT     MORTGAGE OWN      ANY
Levels: ANY MORTGAGE OWN RENT

> loan$home_ownership <- as.character(loan$home_ownership)

> loan$home_ownership[loan$home_ownership=="OWN" | loan$home_ownership=="MORT
GAGE"  ] <- 1

> loan$home_ownership[loan$home_ownership!=1] <- 0

> loan$home_ownership <- as.numeric(loan$home_ownership)

> unique(loan$home_ownership)
[1] 0 1
```

Binarization of purpose

Purpose variable was binarize based on Lending Club offer and intuition. As one of these values refers to personal needs and the other parts to financial issues. I decided to binarize this variable as shown in the below code.

```
> unique(loan$purpose)
 [1] credit_card        other              house              debt_consolidat
ion
 [5] car                home_improvement   moving             major_purchase
 [9] vacation           small_business     medical            renewable_energ
y
14 Levels: car credit_card debt_consolidation educational home_improvement ..
. wedding

> loan$purpose <- as.character(loan$purpose)

> loan$purpose[loan$purpose == "home_improvement" | loan$purpose == "other" |
loan$purpose == "moving" | loan$purpose == "vacation" |
+                loan$purpose == "major_purchase"| loan$purpose == "small_bus
iness"| loan$purpose == "car" | loan$purpose == "medical"|
+                loan$purpose == "house" | loan$purpose == "renewable_energy"
| loan$purpose == "wedding"] <- 1

> loan$purpose[loan$purpose != 1] <- 0

> loan$purpose <- as.numeric(loan$purpose)

> unique(loan$purpose)
[1] 0 1
```

Clearance of issue_d

```
> head(loan$issue_d)
[1] Dec-2018 Dec-2018 Dec-2018 Dec-2018 Dec-2018 Dec-2018
36 Levels: Apr-2015 Apr-2016 Apr-2018 Aug-2015 Aug-2016 Aug-2018 ... Sep-2018
```

```
> loan$issue_d <- as.character(loan$issue_d)
```

```
> substrRight <- function(x, n){
+   substr(x, nchar(x)-n+1, nchar(x))
+ }
```

```
> loan$issue_d <- substrRight(loan$issue_d, 4)
```

```
> loan$issue_d <- as.numeric(loan$issue_d)
```

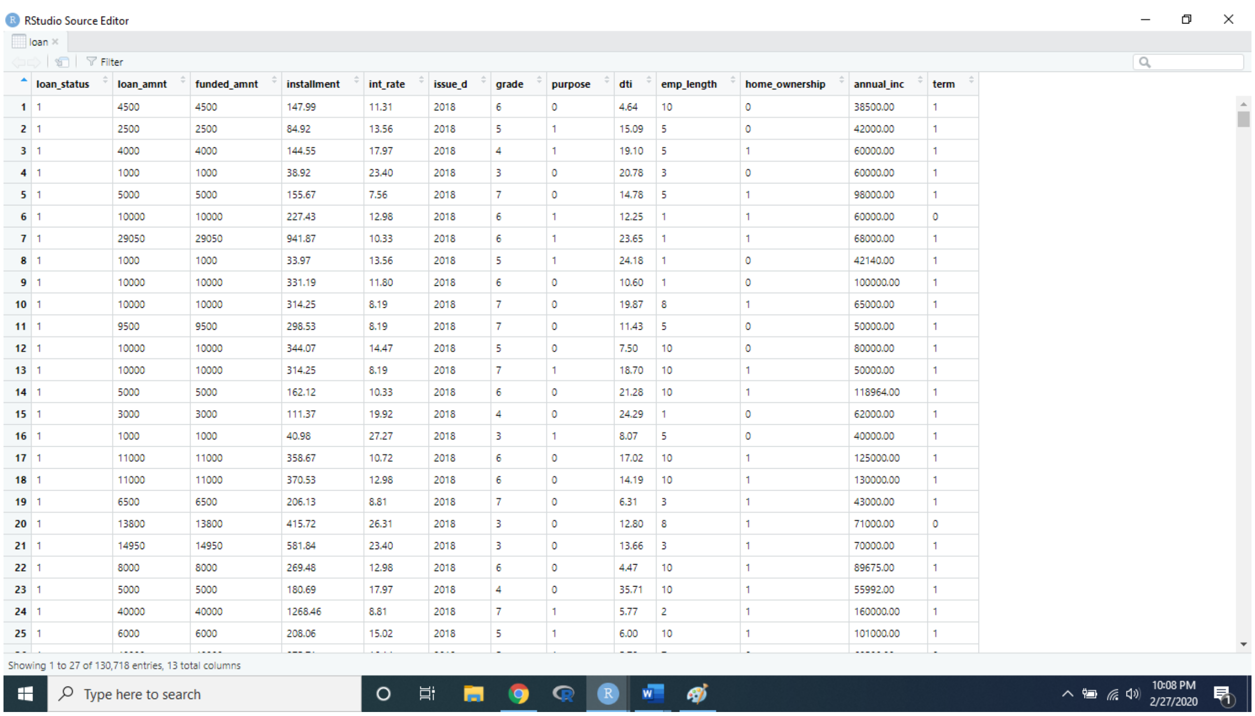
Binarization of dependent variable loan_status

```
> loan$loan_status <- as.character(loan$loan_status)
```

```
> loan$loan_status[loan$loan_status == "Fully Paid"] <- 1
```

```
> loan$loan_status[loan$loan_status != 1] <- 0
```

```
> loan$loan_status <- as.numeric(loan$loan_status)
```

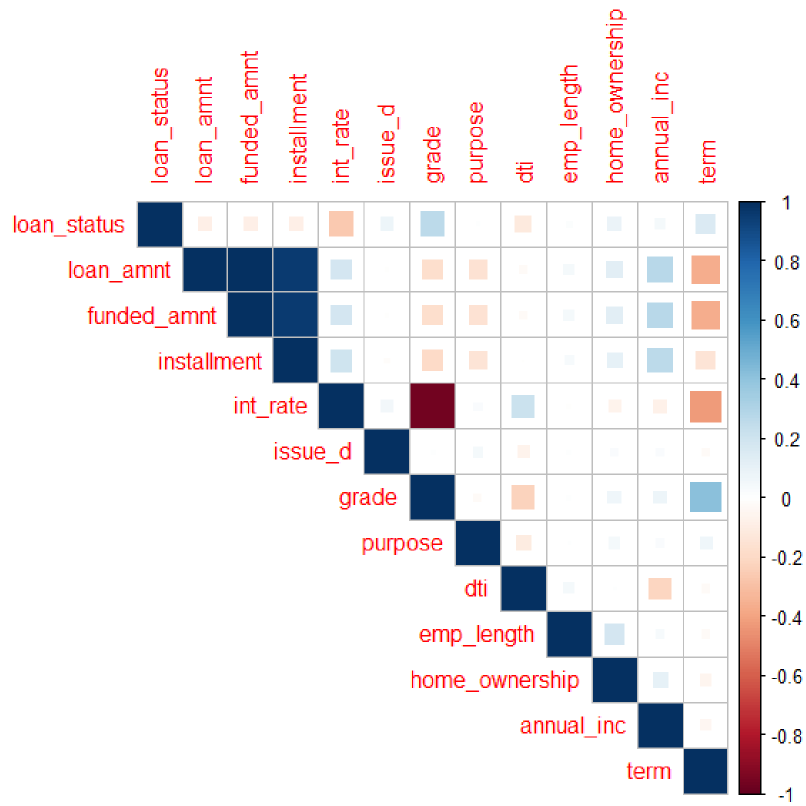


	loan_status	loan_amnt	funded_amnt	installment	int_rate	issue_d	grade	purpose	dti	emp_length	home_ownership	annual_inc	term
1	1	4500	4500	147.99	11.31	2018	6	0	4.64	10	0	38500.00	1
2	1	2500	2500	84.92	13.56	2018	5	1	15.09	5	0	42000.00	1
3	1	4000	4000	144.55	17.97	2018	4	1	19.10	5	1	60000.00	1
4	1	1000	1000	38.92	23.40	2018	3	0	20.78	3	0	60000.00	1
5	1	5000	5000	155.67	7.56	2018	7	0	14.78	5	1	98000.00	1
6	1	10000	10000	227.43	12.98	2018	6	1	12.25	1	1	60000.00	0
7	1	29050	29050	941.87	10.33	2018	6	1	23.65	1	1	68000.00	1
8	1	1000	1000	33.97	13.56	2018	5	1	24.18	1	0	42140.00	1
9	1	10000	10000	331.19	11.80	2018	6	0	10.60	1	0	100000.00	1
10	1	10000	10000	314.25	8.19	2018	7	0	19.87	8	1	65000.00	1
11	1	9500	9500	298.53	8.19	2018	7	0	11.43	5	0	50000.00	1
12	1	10000	10000	344.07	14.47	2018	5	0	7.50	10	0	80000.00	1
13	1	10000	10000	314.25	8.19	2018	7	1	18.70	10	1	50000.00	1
14	1	5000	5000	162.12	10.33	2018	6	0	21.28	10	1	118964.00	1
15	1	3000	3000	111.37	19.92	2018	4	0	24.29	1	0	62000.00	1
16	1	1000	1000	40.98	27.27	2018	3	1	8.07	5	0	40000.00	1
17	1	11000	11000	358.67	10.72	2018	6	0	17.02	10	1	125000.00	1
18	1	11000	11000	370.53	12.98	2018	6	0	14.19	10	1	130000.00	1
19	1	6500	6500	206.13	8.81	2018	7	0	6.31	3	1	43000.00	1
20	1	13800	13800	415.72	26.31	2018	3	0	12.80	8	1	71000.00	0
21	1	14950	14950	581.84	23.40	2018	3	0	13.66	3	1	70000.00	1
22	1	8000	8000	269.48	12.98	2018	6	0	4.47	10	1	89675.00	1
23	1	5000	5000	180.69	17.97	2018	4	0	35.71	10	1	55992.00	1
24	1	40000	40000	1268.46	8.81	2018	7	1	5.77	2	1	160000.00	1
25	1	6000	6000	208.06	15.02	2018	5	1	6.00	10	1	101000.00	1

Conclusion – As we can see the entire data is now categorized.

Looking for correlation

```
> Corr_ <- cor(loan)
> corrplot(Corr_, method = "square", type = "upper")
```



Conclusion - We can see that loan_amnt, funded_amnt and installment are highly positively correlated. Also int_rate and grade are highly negatively correlated the refore we remove funded_amnt, installment
and grade.

PCA Application

```
> ABC<- loan[,c("loan_amnt" , "int_rate", "issue_d" , "purpose", "dti",
+               "emp_length" , "home_ownership" ,"annual_inc" , "term")]
```

```
> cor(ABC[,-1])
```

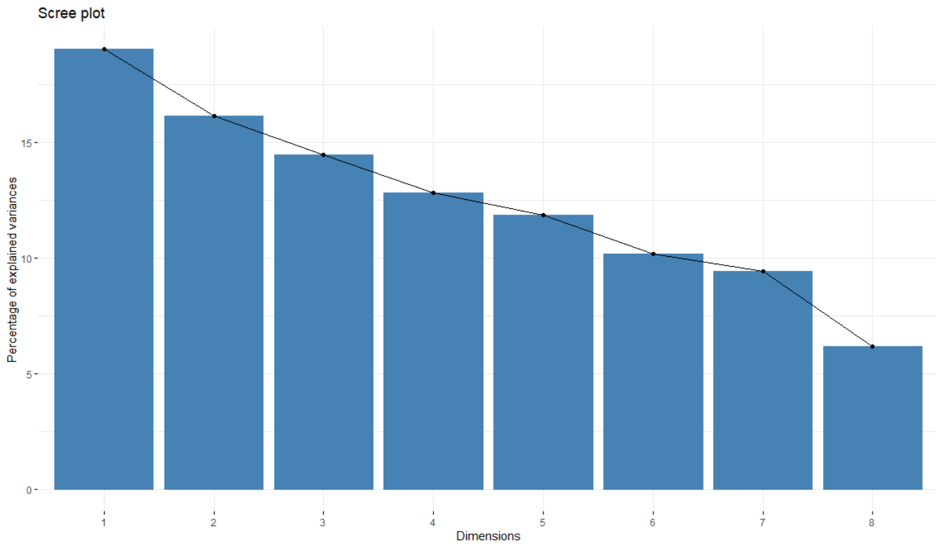
```
+                "emp_length" , "home_ownership" ,"annual_inc" , "term")]
> cor(ABC[,-1])
                  int_rate     issue_d     purpose          dti   emp_length
int_rate        1.00000000  0.056567247  0.026051583  0.216011067 -0.007511820
issue_d         0.05656725  1.000000000  0.046423193 -0.067248173  0.009880283
purpose         0.02605158  0.046423193  1.000000000 -0.100760314  0.005532303
dti             0.21601107 -0.067248173 -0.100760314  1.000000000  0.048241195
emp_length     -0.00751182  0.009880283  0.005532303  0.048241195  1.000000000
home_ownership -0.06675641  0.023826468  0.048543854  0.004595834  0.181587513
annual_inc     -0.07140994  0.024494141  0.026711033 -0.213059622  0.030683145
term           -0.42675736 -0.023469236  0.063222698 -0.026617172 -0.029370193
               home_ownership  annual_inc        term
int_rate         -0.066756409 -0.07140994 -0.42675736
issue_d           0.023826468  0.02449414 -0.02346924
purpose           0.048543854  0.02671103  0.06322270
dti               0.004595834 -0.21305962 -0.02661717
emp_length        0.181587513  0.03068314 -0.02937019
home_ownership    1.000000000  0.10196960 -0.05784198
annual_inc        0.101969604  1.00000000 -0.04585242
term             -0.057841985 -0.04585242  1.00000000
>
```

```
> ABC.pca = prcomp(ABC[,-1], scale. = TRUE)
```

```
> summary(ABC.pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7     P
C8
Standard deviation     1.2337 1.1353 1.0747 1.0125 0.9736 0.9015 0.8681 0.703
51
Proportion of Variance 0.1902 0.1611 0.1444 0.1281 0.1185 0.1016 0.0942 0.061
87
Cumulative Proportion  0.1902 0.3514 0.4957 0.6239 0.7423 0.8439 0.9381 1.000
00
```

```
> fviz_eig(ABC.pca)
```



Conclusion – From the above graph we decide to include 5 pcas's as these components will help us maximize the total variance.