PES University, Bengaluru

UE18CS312 – Data Analytics

Worksheet 1 (for Unit 1)

Submitted by: Venkatavaradan R SRN: PES2201800462 Branch: CSE Section: E

1. Loading the dataset

```
bkb<-read.csv(here::here("data","BKB.csv"),stringsAsFactors = FALSE)
bkb<-tibble(bkb)
```

2. Summary of the data

```
bkb %>% summary()
```

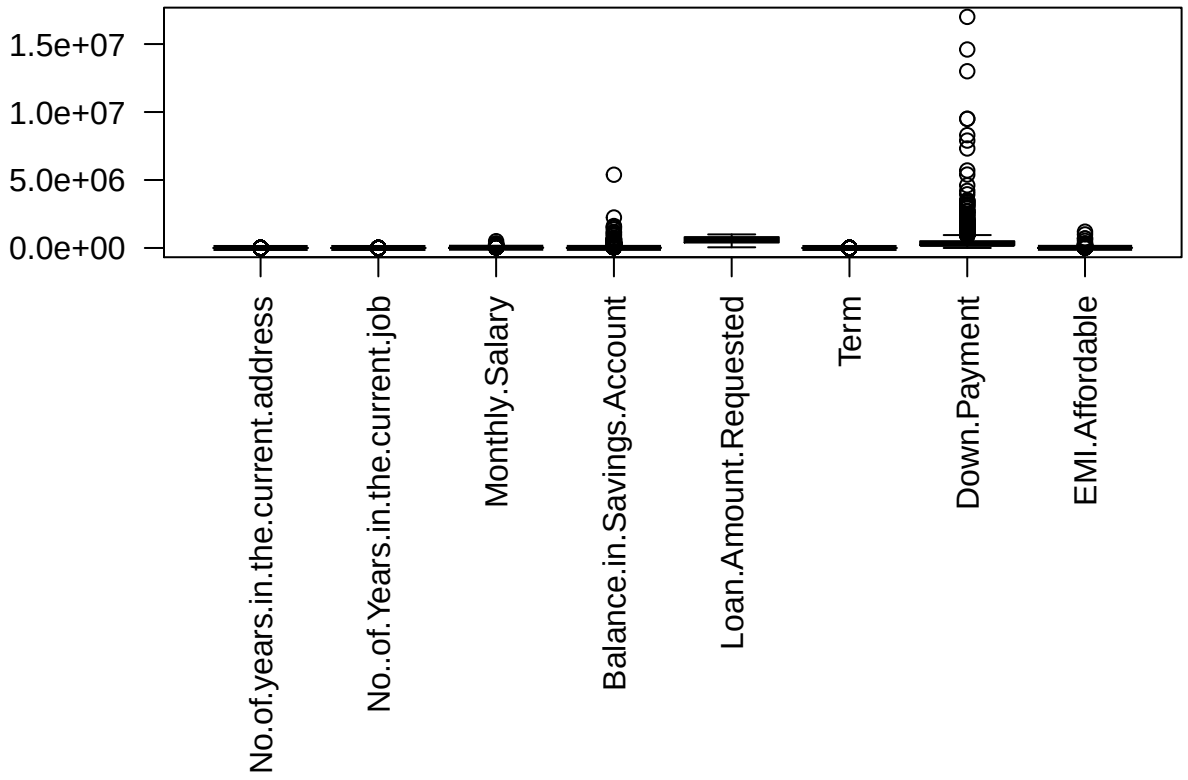
```
##   Applicant.ID     Loan.Type            Gender          Marital.Status
##  Min.   :   1.0   Length:3864        Length:3864        Length:3864
##  1st Qu.: 966.8   Class :character   Class :character   Class :character
##  Median :1932.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1932.5
##  3rd Qu.:2898.2
##  Max.   :3864.0
##  Accomodation.Type  No.of.years.in.the.current.address
##  Length:3864        Min.   : 0.0
##  Class :character   1st Qu.: 2.0
##  Mode  :character   Median : 6.0
##                     Mean   :10.6
##                     3rd Qu.:15.0
##                     Max.   :92.0
##  No..of.Years.in.the.current.job Monthly.Salary   Balance.in.Savings.Account
##  Min.   : 0.00                   Min.   :     0   Min.   :      0
##  1st Qu.: 5.00                   1st Qu.: 12201   1st Qu.:   1500
##  Median :10.00                   Median : 19000   Median :   6358
##  Mean   :10.93                   Mean   : 22619   Mean   :  31583
##  3rd Qu.:15.00                   3rd Qu.: 28500   3rd Qu.:  25000
##  Max.   :65.00                   Max.   :500000   Max.   :5388413
##  Loan.Amount.Requested      Term        Down.Payment      EMI.Affordable
##  Min.   :  50000       Min.   : 15.0   Min.   :       0   Min.   :     84
##  1st Qu.: 400000       1st Qu.:180.0   1st Qu.:  200000   1st Qu.:   7696
##  Median : 600000       Median :180.0   Median :  300000   Median :  10774
##  Mean   : 609055       Mean   :160.2   Mean   :  427471   Mean   :  12882
##  3rd Qu.: 800000       3rd Qu.:180.0   3rd Qu.:  500000   3rd Qu.:  15000
##  Max.   :1000000       Max.   :180.0   Max.   :17000000   Max.   :1200000
```

```
bkb_num<-bkb %>%
  select_if(is.numeric)%>%
  select(-Applicant.ID)
```

Creating numeric df

3. Boxplot

```
bkb_num %>%
  boxplot(las=2)
```



We can clearly see that there are many outliers present for savings account, down payment and emi affordable. Lets dive deeper and find out the count.

```
bkb_num %>%
     select(Balance.in.Savings.Account,Down.Payment,EMI.Affordable) %>%
      map(~ boxplot.stats(.x)$out)%>%
      map(~ length(.x))
```

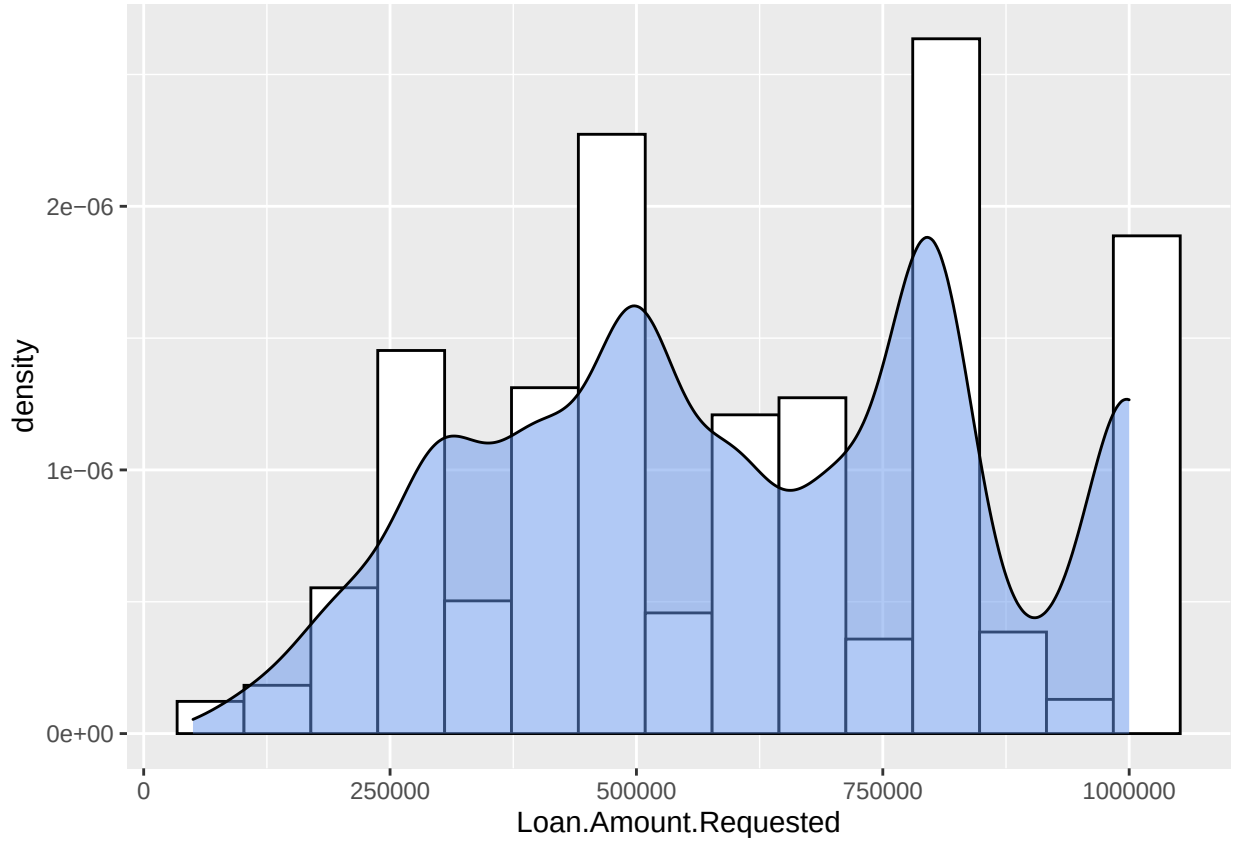
finding count of outliers

```
## $Balance.in.Savings.Account
## [1] 417
##
## $Down.Payment
## [1] 274
##
## $EMI.Affordable
## [1] 45
```

4. Histogram + density plot

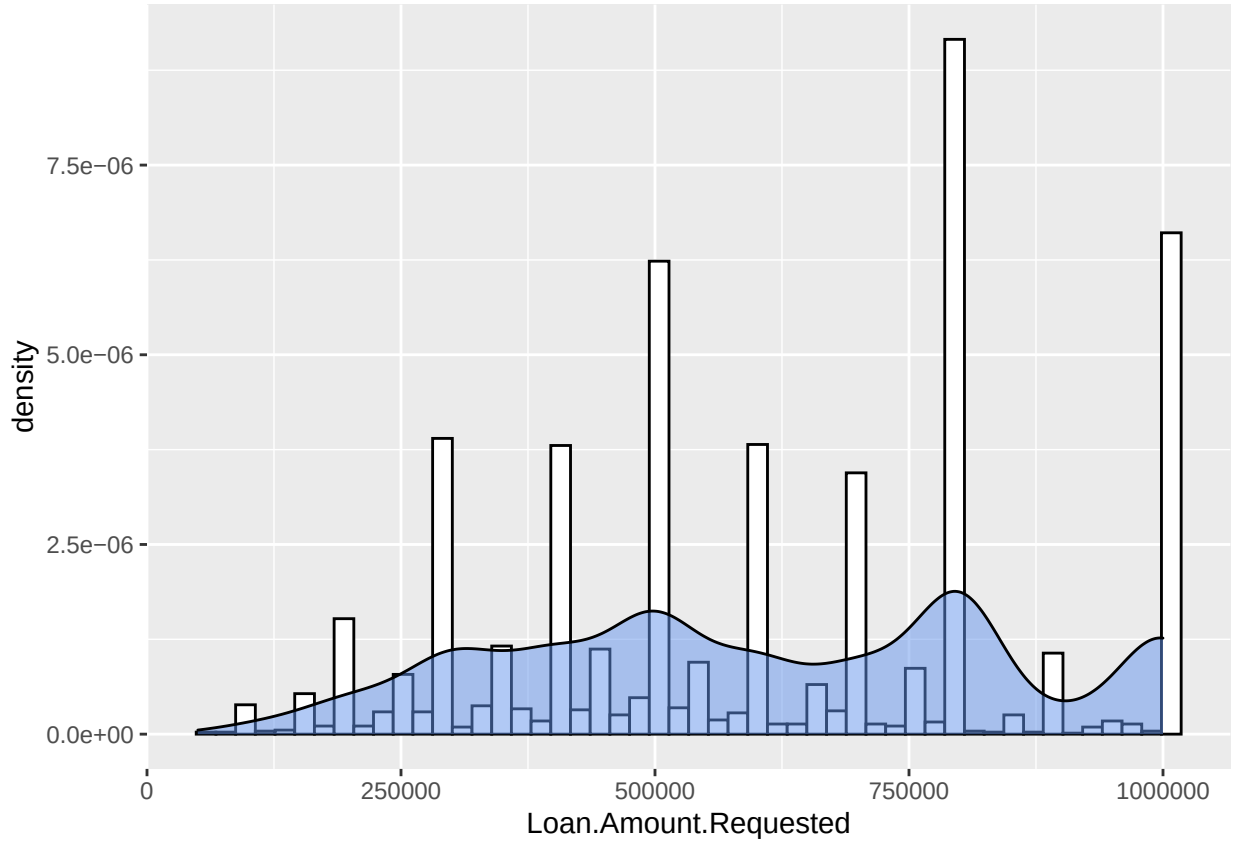
(bin width = 15)

```
bkb_num %>%
  ggplot(aes(Loan.Amount.Requested))+
  geom_histogram(aes(y=..density..),bins=15,fill="white",color="black")+
  geom_density(fill = "cornflowerblue", alpha=0.5)
```



(bin width = 50)

```
bkb_num %>%
  ggplot(aes(Loan.Amount.Requested))+
  geom_histogram(aes(y=..density..),bins=50,fill="white",color="black")+
  geom_density(fill = "cornflowerblue", alpha=0.5)
```



A barchart will be suitable for visualization of Loan Amount Variable

5. Hypothesis Testing

Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins at the same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

H0 : Mean weight of penguins do not differ from last year. HA : Mean weight of penguins differs from last year.

Since no of observations > 30, we will use Z-test instead of T-test.

```
st_err = 2.5/(35**0.5)
z = (14.6-15.4)/st_err
pnorm(z)*2
```

```
## [1] 0.05833852
```

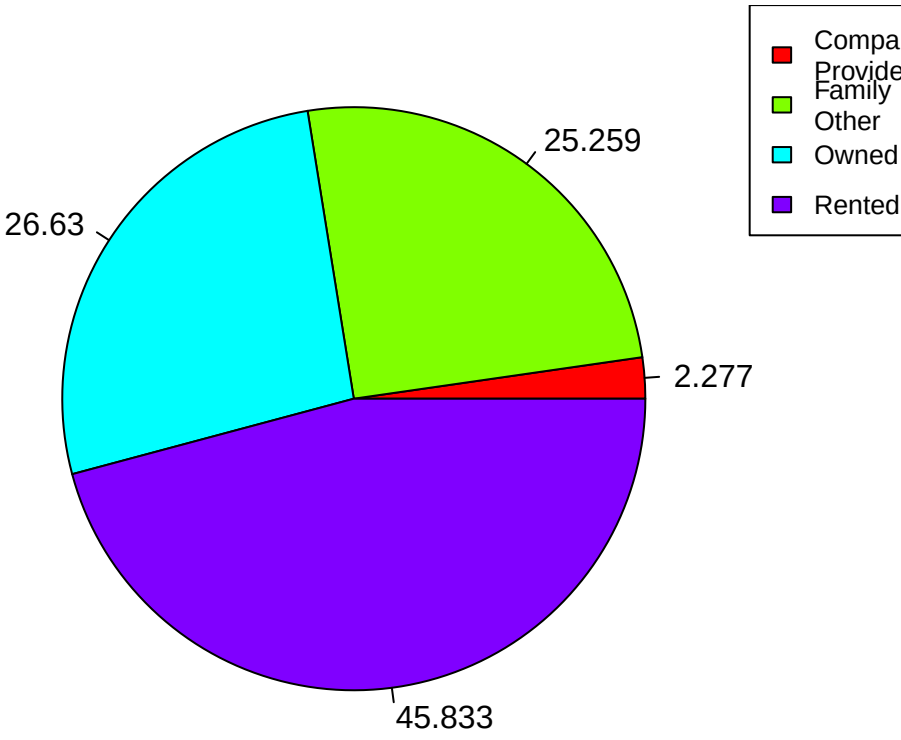
Since p>0.05, we reject H0 , thus at 0.05 significance level we reject null hypothesis and accept the alternate hypothesis that Mean weight of penguins differ from last year.

6. Accomodation type Visualization - Pie

```
accomodation_details<-bkb%>%
  select(Accomodation.Type)%>%
  group_by(Accomodation.Type)%>%
  count()
```

Creating a tibble with accomodation name and count grouped by Accomodation Type

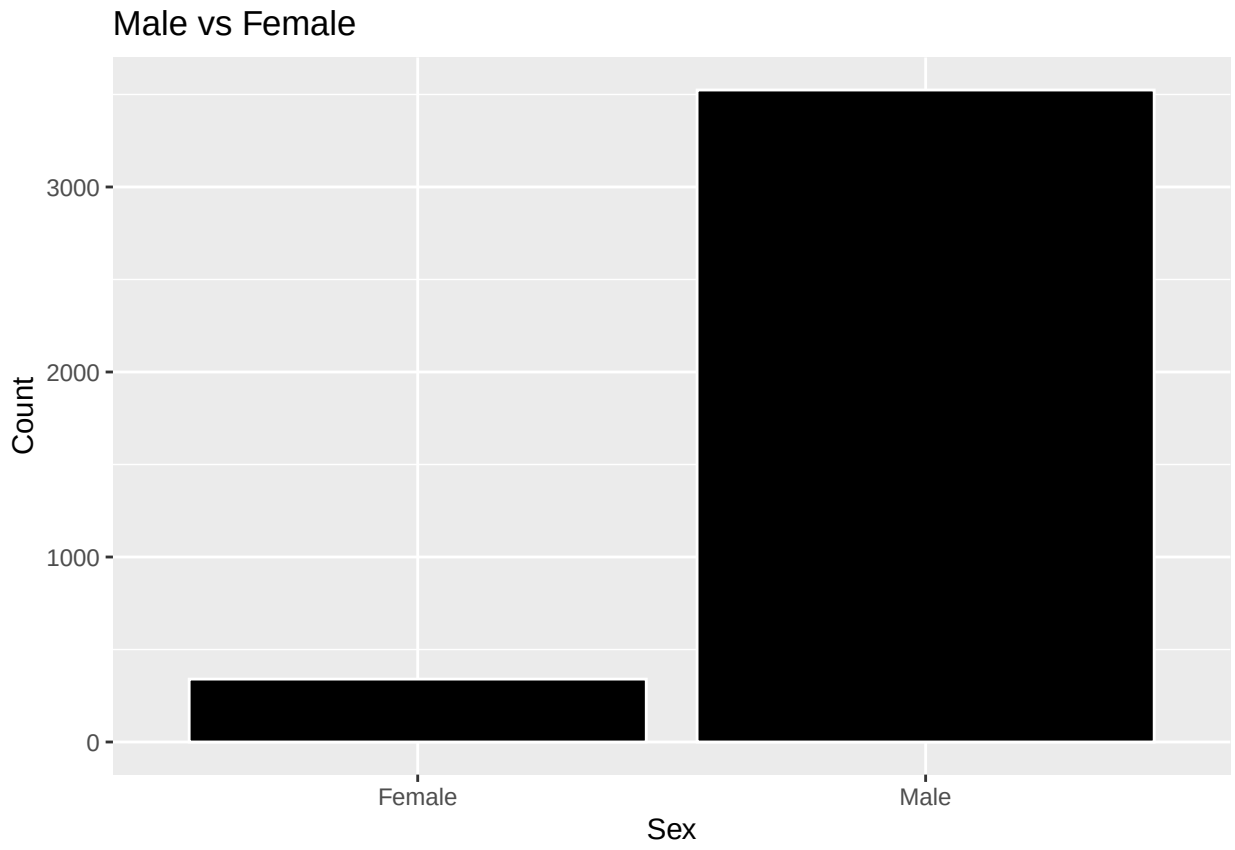
```
t <- table(bkb$Accomodation.Type)
piepercent <- t*100/(sum(t))
pie(table(bkb$Accomodation.Type),
labels=round(piepercent,digits=3),col = rainbow(length(t)))
legend("topright", c("Company
Provided","Family
Other","Owned","Rented"), cex= 0.8, fill =
rainbow(length(t)))
```



Plotting the graph

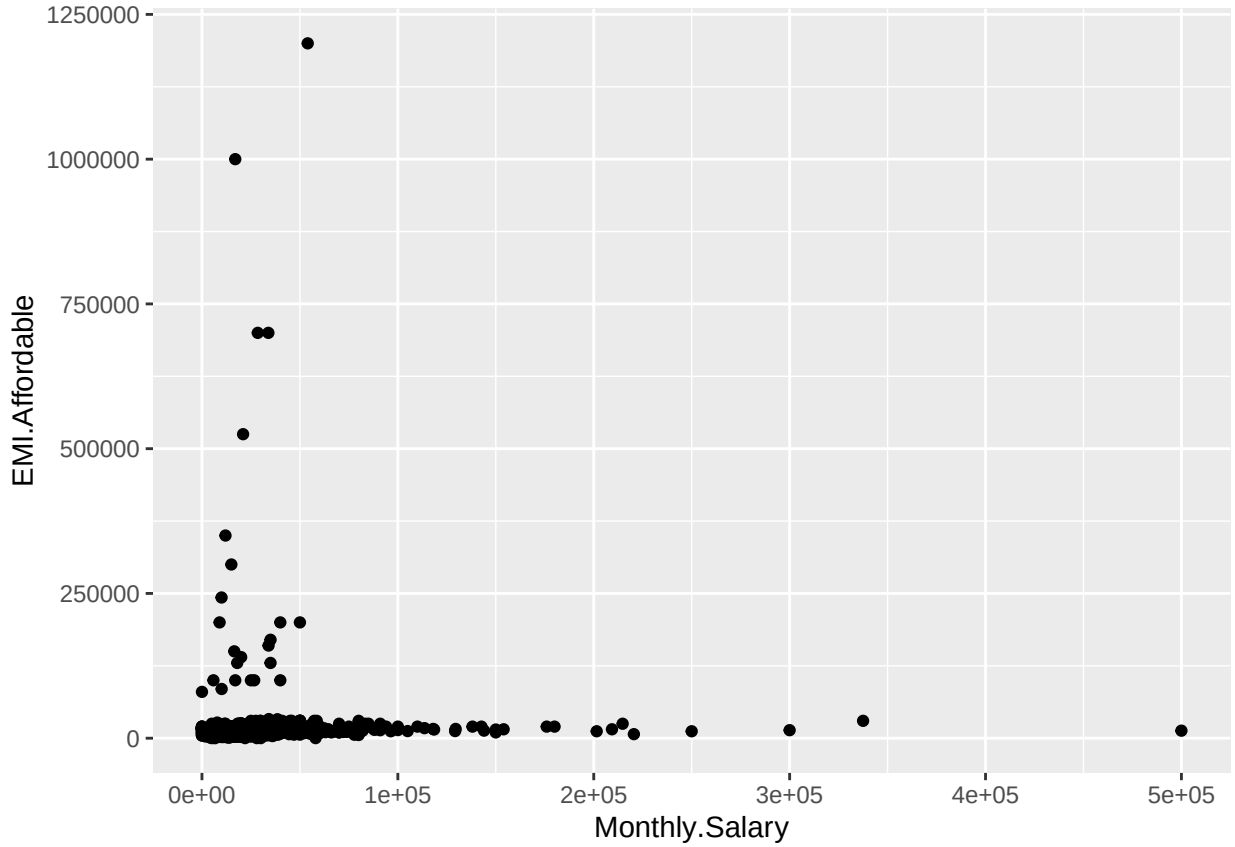
7. Gender type Visualization - Bar

```
bkb%>%
  ggplot(aes(Gender))+
  geom_bar(fill="black",color="white")+
  ggtitle("Male vs Female")+
  xlab("Sex")+
  ylab("Count")
```



8. Monthly Salaries wrt EMI - scatter

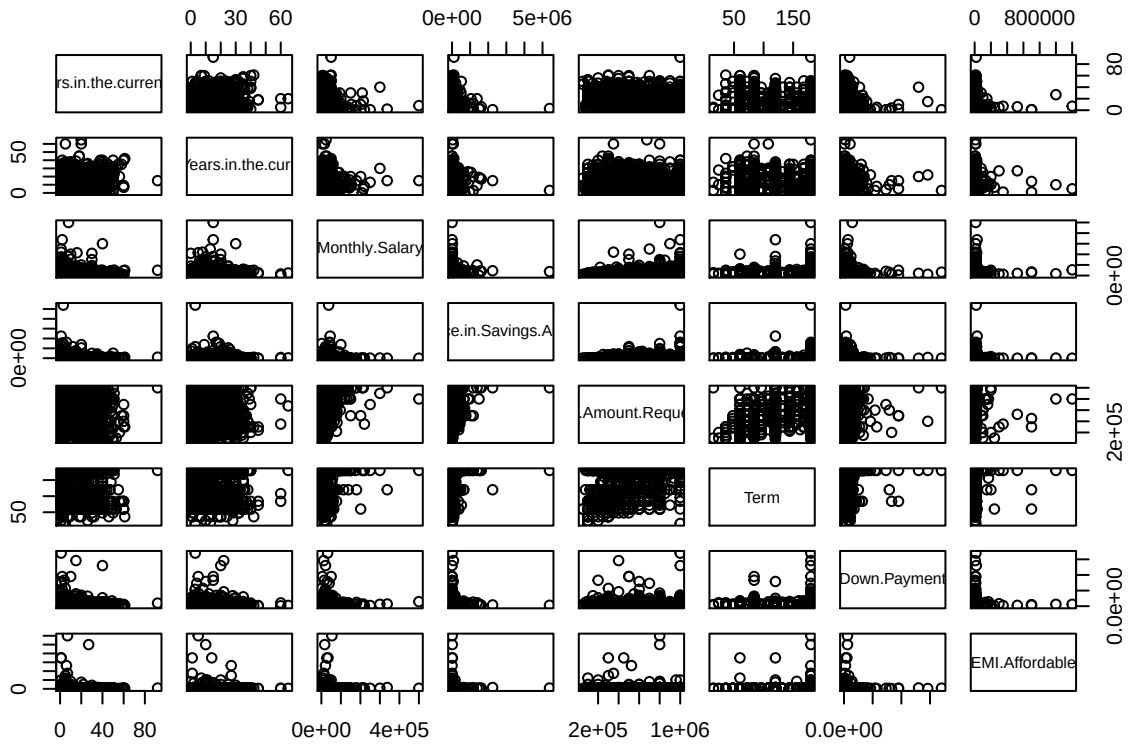
```
bkb_num%>%
  ggplot(aes(Monthly.Salary,EMI.Affordable))+
  geom_point()
```



When the Monthly salary is less, the Loan Amount is also less and is perfectly manageable to pay the EMIs. There are also outliers that lie in both directions where either EMI is too high or Salary is way higher when compared to EMI rate.

```
# bkb_num%>%
#   ggpairs()

pairs(bkb_num)
```



pairplots

9. Descriptive statistics for Salary

```
mean(bkb_num$Monthly.Salary)
```

```
## [1] 22618.98
```

```
median(bkb_num$Monthly.Salary)
```

```
## [1] 19000
```

```
range(bkb_num$Monthly.Salary)
```

```
## [1]      0 500000
```

```
sd(bkb_num$Monthly.Salary)
```

```
## [1] 19783.32
```

```
bkb_num%>%
  mutate(
    Monthly.Salary = as.factor(Monthly.Salary)
  )%>%
  select(Monthly.Salary)%>%
  group_by(Monthly.Salary)%>%
  count()%>%
  ungroup()%>%
  top_n(1)
```

```
## Selecting by freq
```

```
##   Monthly.Salary freq
## 1          15000  233
```

10. Salary analytics

```
q10a<-bkb%>%
  select(Monthly.Salary,Gender)%>%
  filter(Gender=="Female")
mean(q10a$Monthly.Salary)
```

mean monthly salary for females

```
## [1] 19675.38
```

```
q10b<-bkb%>%
  select(Monthly.Salary,Gender)%>%
  filter(Gender=="Male")
median(q10b$Monthly.Salary)
```

median monthly salary for males

```
## [1] 19479.5
```

11. Monthly salaries grouped by the Gender

```
bkb%>%
  select(Monthly.Salary,Gender)%>%
  group_by(Gender)%>%
  summarize_at(vars(Monthly.Salary),list(mean))
```

mean

```
## # A tibble: 2 x 2
##   Gender Monthly.Salary
##   <chr>           <dbl>
## 1 Female         19675.
## 2 Male           22903.
```

```
bkb%>%
  select(Monthly.Salary,Gender)%>%
  group_by(Gender)%>%
  summarize_at(vars(Monthly.Salary),list(median))
```

median

```
## # A tibble: 2 x 2
##   Gender Monthly.Salary
##   <chr>           <dbl>
## 1 Female         15486.
## 2 Male           19480.
```

```
bkb%>%
  select(Monthly.Salary,Gender)%>%
  group_by(Gender)%>%
  summarize_at(vars(Monthly.Salary),list(range))
```

range

```
## # A tibble: 4 x 2
## # Groups:   Gender [2]
##   Gender Monthly.Salary
##   <chr>           <int>
## 1 Female              0
## 2 Female         110000
## 3 Male                0
## 4 Male           500000
```

12. Skewness and kurtosis for Monthly Salary

```
skewness(bkb$Monthly.Salary)
```

```
## [1] 7.950902
```

```
kurtosis(bkb$Monthly.Salary)
```

```
## [1] 134.3941
```

As observed from the histogram above and the values of skewness and kurtosis, we infer that Monthly Salary is Positively Skewed(Rightly Skewed) and LeptoKurtic -> Highly centered peak and has small tails.

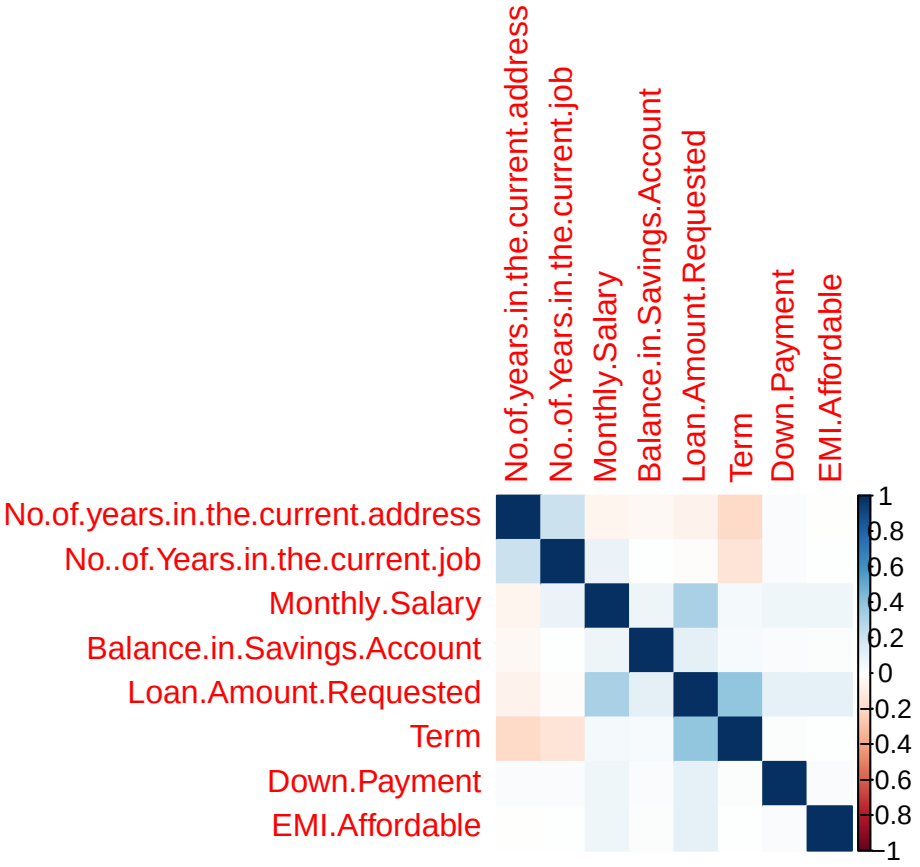
13. Correlation between Loan amount and Down payment

```
bkb_num%>%
  select(Loan.Amount.Requested,Down.Payment)%>%
  cor()
```

```
##                          Loan.Amount.Requested Down.Payment
## Loan.Amount.Requested                1.0000000    0.1055291
## Down.Payment                         0.1055291    1.0000000
```

14. Correlogram

```
c <- cor(bkb_num)
corrplot(c,method = 'color')
```



15. PCA

```
pc <- prcomp(bkb_num,center = TRUE, scale = TRUE)

pc%>%
  summary()
```

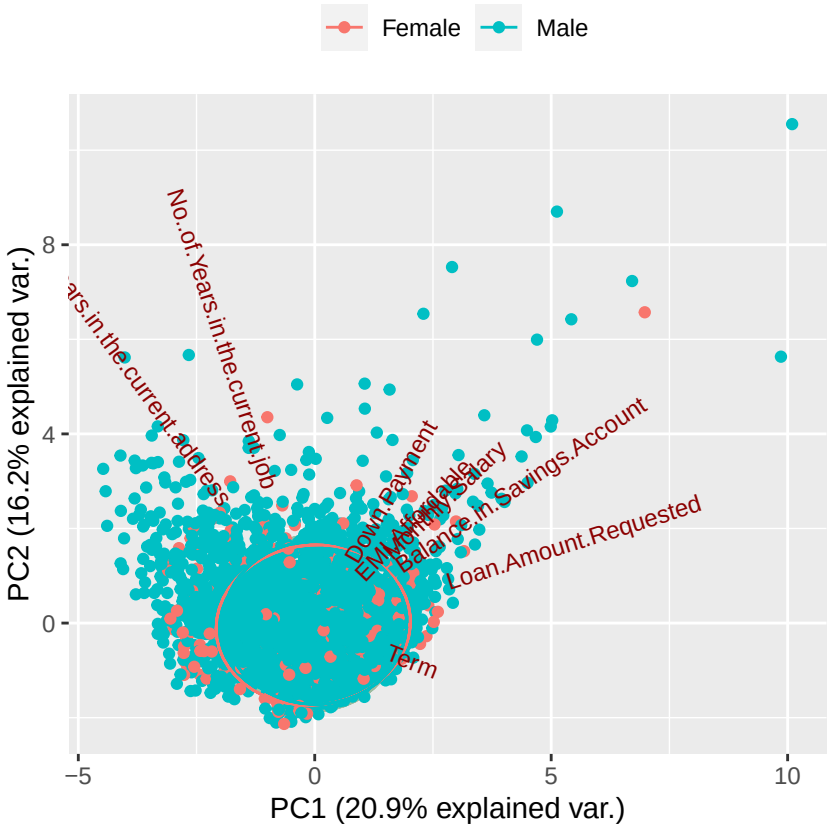
```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6     PC7
## Standard deviation     1.2921 1.1398 0.9952 0.9881 0.9752 0.9320 0.86866
## Proportion of Variance 0.2087 0.1624 0.1238 0.1221 0.1189 0.1086 0.09432
## Cumulative Proportion  0.2087 0.3711 0.4949 0.6169 0.7358 0.8444 0.93869
##                           PC8
## Standard deviation     0.70032
## Proportion of Variance 0.06131
## Cumulative Proportion  1.00000
```

```
predict(pc)%>%
  round(2)%>%
  head()
```

```
##        PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
## [1,] -1.64  0.56  0.10  0.59  0.44  0.80  1.39 -0.25
## [2,]  1.48  0.47 -0.06 -0.20 -0.66  0.05 -0.75 -0.78
## [3,] -0.35 -0.14  0.20  0.56  0.93 -0.23  0.01 -1.15
## [4,] -2.11 -0.11 -0.02 -0.45  0.60 -0.99  0.82 -0.13
## [5,] -0.22 -0.81 -0.04  0.54  0.32 -0.07 -0.62  0.62
## [6,] -0.07 -0.04 -0.02  0.65  0.20  1.01  1.16  0.43
```

16. PCA Visualization

```
pc <- prcomp(bkb_num,center=TRUE, scale. = TRUE)
ggbiplot(pc, obs.scale = 1, var.scale = 1, groups = bkb$Gender, ellipse = TRUE, circle = TRUE)+ scale_c
theme(legend.direction = 'horizontal', legend.position = 'top')
```


Female
Male
PC2 (16.2% explained var.)
PC1 (20.9% explained var.)
8
4
0
−5
0
5
10
ars.in.the.current.address
No..of.Years.in.the.current.job
Down.Payment
EMI.Affordable
Monthly.Salary
Balance.in.Savings.Account
Loan.Amount.Requested
Term