Data manipulation(Rprogramming)

SAGAR MEHTA

14/05/2020

1.Load the required libraries and the data.

```
housing_data <- read.csv('housingdata_v2.0.csv')
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

2.Understand the data structure and provide concise summary on the following –•no of observations•total number of variables•number of continuous variables•number of categorical variables

```
head(housing_data)
```

```
##     Record Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0     Graduate     No    710         90400         456
## 2  Record8   Male       0     Graduate     No   6516        168800         336
## 3  Record9   Male       0     Graduate    Yes   7040        160000         336
## 4 Record10   Male       0 Not Graduate     No   4730        155200         336
## 5 Record11   Male       0     Graduate     No   9167        149600         336
## 6 Record12   Male       0     Graduate     No  10459        149600         336
##   Credit_Record Housing_type Property_Purchased
## 1             1   Affordable                  Y
## 2             1   Affordable                  Y
## 3             1   Affordable                  Y
## 4             1   Affordable                  Y
## 5             1   Affordable                  Y
## 6             1   Affordable                  Y
```

```
nrow(housing_data)
```

```
## [1] 505
```

```
ncol(housing_data)
```

```
## [1] 11
```

```
dim(housing_data)
```

```
## [1] 505  11
```

```
class(housing_data)
```

```
## [1] "data.frame"
```

```
str(housing_data)
```

```
## 'data.frame':    505 obs. of  11 variables:
##  $ Record           : Factor w/ 505 levels "Record1","Record10",..: 1 484 495 2 13 24 35 57 68 90 ...
##  $ Gender           : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 2 2 2 ...
##  $ No_kids          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 1 1 1 1 ...
##  $ HasCar           : Factor w/ 3 levels "No","Not Answered",..: 1 1 3 1 1 1 1 3 1 1 ...
##  $ Income           : int  710 6516 7040 4730 9167 10459 2888 10960 8692 4044 ...
##  $ PropertyValue    : int  90400 168800 160000 155200 149600 149600 149600 144000 144000 137600 ...
##  $ Loan_Period      : int  456 336 336 336 336 336 336 336 336 336 ...
##  $ Credit_Record    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Housing_type     : Factor w/ 3 levels "Affordable","Mid Range",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Property_Purchased: Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
```

3.Select and Mutate : use the select() and mutate() functions in R to answer the following •Select the columns Gender, Education, and Income and print the first five rows•Select the columns from Gender to Loan Period and print the first five rows•Be concise! -select columns by removing Record Column and Gender and print the first five rows•Use mutate() function to add the new variables var1 which calculates the ratio of property value to total income and save the result as g1. Print the first five rows.•Add the new variable var2 which returns the ratio of property value to loan period and save the result as g2. Print the first five rows.

#Select the columns Gender, Education, and Income and print the first five rows

```
housing_gei <- housing_data%>%select(Gender,Education,Income)
head(housing_gei)
```

```
##   Gender    Education Income
## 1 Female     Graduate    710
## 2   Male     Graduate   6516
## 3   Male     Graduate   7040
## 4   Male Not Graduate   4730
## 5   Male     Graduate   9167
## 6   Male     Graduate  10459
```

#Select the columns from Gender to Loan Period and print the first five rows

```
housing_glp <- housing_data%>%select(Gender:Loan_Period)
head(housing_glp)
```

```
##   Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1 Female       0     Graduate     No    710         90400         456
## 2   Male       0     Graduate     No   6516        168800         336
## 3   Male       0     Graduate    Yes   7040        160000         336
## 4   Male       0 Not Graduate     No   4730        155200         336
## 5   Male       0     Graduate     No   9167        149600         336
## 6   Male       0     Graduate     No  10459        149600         336
```

#Be concise! -select columns by removing Record Column and Gender and print the first five rows

```
housing_rcg <- housing_data%>%select(3:11)
head(housing_rcg)
```

```
##   No_kids    Education HasCar Income PropertyValue Loan_Period Credit_Record
## 1       0     Graduate     No    710         90400         456             1
## 2       0     Graduate     No   6516        168800         336             1
## 3       0     Graduate    Yes   7040        160000         336             1
## 4       0 Not Graduate     No   4730        155200         336             1
## 5       0     Graduate     No   9167        149600         336             1
## 6       0     Graduate     No  10459        149600         336             1
##   Housing_type Property_Purchased
## 1   Affordable                  Y
## 2   Affordable                  Y
## 3   Affordable                  Y
## 4   Affordable                  Y
## 5   Affordable                  Y
## 6   Affordable                  Y
```

#Use mutate() function to add the new variables var1 which calculates the ratio of property value to total income and save the result as g1. Print the first five rows.

```
g1 <- housing_data%>%mutate(var1 = PropertyValue/Income)
head(g1)
```

```
##     Record Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0     Graduate     No    710         90400         456
## 2  Record8   Male       0     Graduate     No   6516        168800         336
## 3  Record9   Male       0     Graduate    Yes   7040        160000         336
## 4 Record10   Male       0 Not Graduate     No   4730        155200         336
## 5 Record11   Male       0     Graduate     No   9167        149600         336
## 6 Record12   Male       0     Graduate     No  10459        149600         336
##   Credit_Record Housing_type Property_Purchased      var1
## 1             1   Affordable                  Y 127.32394
## 2             1   Affordable                  Y  25.90546
## 3             1   Affordable                  Y  22.72727
## 4             1   Affordable                  Y  32.81184
## 5             1   Affordable                  Y  16.31941
## 6             1   Affordable                  Y  14.30347
```

#Add the new variable var2 which returns the ratio of property value to loan period and save the result as g2. Print the first five rows.

```
g2 <- housing_data%>%mutate(var2 = PropertyValue/Loan_Period)
head(g2)
```

```
##     Record Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0     Graduate     No    710         90400         456
## 2  Record8   Male       0     Graduate     No   6516        168800         336
## 3  Record9   Male       0     Graduate    Yes   7040        160000         336
## 4 Record10   Male       0 Not Graduate     No   4730        155200         336
## 5 Record11   Male       0     Graduate     No   9167        149600         336
## 6 Record12   Male       0     Graduate     No  10459        149600         336
##   Credit_Record Housing_type Property_Purchased     var2
## 1             1   Affordable                  Y 198.2456
## 2             1   Affordable                  Y 502.3810
## 3             1   Affordable                  Y 476.1905
## 4             1   Affordable                  Y 461.9048
## 5             1   Affordable                  Y 445.2381
## 6             1   Affordable                  Y 445.2381
```

4.Filter and Arrange: •Filter all the observations that have Property Value lower than 80000 or higher than 150000 and store it in df g3. Print the first five rows. How many observations are there.•Filter all the observations that have Property Value > 1000000 and Income < 3185 and store it in df g4. Print the first five rows. How many observations are there.•Filter all observations where Income< 3185 and still Property was purchased. How many such records are there in the data set. Print the first five rows.Use the arrange() function in dplyr to -: •Create a data frame by the name 'bought' –which includes observations when the Property was purchased. How many observations are there.•Arrange the data frame bought by Income and print the first five rows.•Arrange the data frame bought by Gender and print the first five rows.•Arrange the data frame bought so that Gender and Education is grouped and print the first five rows.•Create a data frame by the name 'notbought' –which includes observations when the Property was not purchased. How many observations are there.•Arrange the data frame notbought by Income and print the first five rows.•Arrange the data frame notbought by Gender and print the first five rows.•Arrange the data frame notbought so that Gender and Education is grouped and print the first five rows.•Reverse the order of arranging -Arrange the housing data according to Gender and decreasing Income. Print the first five rows.

Filter all the observations that have Property Value lower than 80000 or higher than 150000 and store it in df g3. Print the first five rows. How many observations are there.

```
g3 <- housing_data%>%filter(PropertyValue<80000|PropertyValue>150000)
head(g3)
```

```
##     Record Gender No_kids    Education       HasCar Income PropertyValue
## 1  Record8    Male       0     Graduate           No   6516        168800
## 2  Record9    Male       0     Graduate          Yes   7040        160000
## 3 Record10    Male       0 Not Graduate           No   4730        155200
## 4 Record76    Male       0 Not Graduate Not Answered   2002         76000
## 5 Record77    Male       0     Graduate           No   3474         71200
## 6 Record78    Male       0     Graduate           No   3212         69600
##   Loan_Period Credit_Record Housing_type Property_Purchased
## 1         336             1   Affordable                  Y
## 2         336             1   Affordable                  Y
## 3         336             1   Affordable                  Y
## 4         336             1   Affordable                  Y
## 5         336             1   Affordable                  Y
## 6         336             1   Affordable                  Y
```

```
dim(g3)
```

```
## [1] 198  11
```

#Filter all the observations that have Property Value > 1000000 and Income < 3185 and store it in df g4. Print the first five rows. How many observations are there

```
g4 <- housing_data%>%filter(PropertyValue > 1000000 & Income < 3185)
head(g4)
```

```
##  [1] Record             Gender             No_kids            Education         
##  [5] HasCar             Income             PropertyValue      Loan_Period       
##  [9] Credit_Record      Housing_type       Property_Purchased
## <0 rows> (or 0-length row.names)
```

```
dim(g4)
```

```
## [1]  0 11
```

#Filter all observations where Income< 3185 and still Property was purchased. How many such records are there in the data set. Print the first five rows.

```
g5 <- housing_data%>%filter(Income < 3185 & Property_Purchased == 'Y')
head(g5)
```

```
##     Record Gender No_kids    Education       HasCar Income PropertyValue
## 1  Record1 Female       0     Graduate           No    710         90400
## 2 Record13   Male       0     Graduate           No   2888        149600
## 3 Record25   Male       0     Graduate           No   3045        124000
## 4 Record26   Male       0 Not Graduate Not Answered   3184        124000
## 5 Record29   Male       0     Graduate          Yes   2835        121600
## 6 Record33   Male       0     Graduate           No   2779        116000
##   Loan_Period Credit_Record Housing_type Property_Purchased
## 1         456             1   Affordable                  Y
## 2         336             1   Affordable                  Y
## 3         336             1   Affordable                  Y
## 4         336             1   Affordable                  Y
## 5         336             1   Affordable                  Y
## 6         336             1   Affordable                  Y
```

```
dim(g5)
```

```
## [1] 81 11
```

#Use the arrange() function in dplyr to -: •Create a data frame by the name 'bought' –which includes observations when the Property was purchased. How many observations are there.

```
bought_property <- housing_data%>%filter(Property_Purchased == 'Y')
head(bought_property)
```

```
##     Record Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0     Graduate     No    710         90400         456
## 2  Record8   Male       0     Graduate     No   6516        168800         336
## 3  Record9   Male       0     Graduate    Yes   7040        160000         336
## 4 Record10   Male       0 Not Graduate     No   4730        155200         336
## 5 Record11   Male       0     Graduate     No   9167        149600         336
## 6 Record12   Male       0     Graduate     No  10459        149600         336
##   Credit_Record Housing_type Property_Purchased
## 1             1   Affordable                  Y
## 2             1   Affordable                  Y
## 3             1   Affordable                  Y
## 4             1   Affordable                  Y
## 5             1   Affordable                  Y
## 6             1   Affordable                  Y
```

#Arrange the data frame bought by Income and print the first five rows.

```
by_income <- bought_property%>%arrange(Income)
head(by_income)
```

```
##      Record Gender No_kids    Education       HasCar Income PropertyValue
## 1 Record202 Female       2 Not Graduate Not Answered    231         78400
## 2   Record1 Female       0     Graduate           No    710         90400
## 3  Record60   Male       0     Graduate           No   1128         89600
## 4 Record313   Male       1     Graduate           No   1788         76800
## 5 Record155   Male       0     Graduate           No   1935        104800
## 6  Record71   Male       1     Graduate           No   1961         85600
##   Loan_Period Credit_Record Housing_type Property_Purchased
## 1         336             1    Mid Range                  Y
## 2         456             1   Affordable                  Y
## 3         336             1   Affordable                  Y
## 4         336             1      Premium                  Y
## 5         336             1    Mid Range                  Y
## 6         336             1   Affordable                  Y
```

#Arrange the data frame bought by Gender and print the first five rows.

```
by_gender <- bought_property%>%arrange(Gender)
head(by_gender)
```

```
##     Record Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0     Graduate     No    710         90400         456
## 2 Record27 Female       0 Not Graduate     No   4785        123200         336
## 3 Record38 Female       0 Not Graduate    Yes   7857        110400         336
## 4 Record41 Female       0     Graduate     No   4139        108000         336
## 5 Record42 Female       0     Graduate     No   5500        105600         336
## 6 Record53 Female       0     Graduate     No   7920         96000         336
##   Credit_Record Housing_type Property_Purchased
## 1             1   Affordable                  Y
## 2             1   Affordable                  Y
## 3             1   Affordable                  Y
## 4             1   Affordable                  Y
## 5             1   Affordable                  Y
## 6             1   Affordable                  Y
```

```
by_education <- bought_property%>%arrange(Gender,Education)
head(by_education)
```

```
##     Record Gender No_kids Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0  Graduate     No    710         90400         456
## 2 Record41 Female       0  Graduate     No   4139        108000         336
## 3 Record42 Female       0  Graduate     No   5500        105600         336
## 4 Record53 Female       0  Graduate     No   7920         96000         336
## 5 Record79 Female       0  Graduate     No   3190         56800         336
## 6 Record81 Female       0  Graduate     No   2758         44800         336
##   Credit_Record Housing_type Property_Purchased
## 1             1   Affordable                  Y
## 2             1   Affordable                  Y
## 3             1   Affordable                  Y
## 4             1   Affordable                  Y
## 5             1   Affordable                  Y
## 6             1   Affordable                  Y
```

#Create a data frame by the name ‘notbought’ –which includes observations when the Property was not purchased. How many observations are there.

```
notbought <- housing_data%>%filter(Property_Purchased == 'N')
head(notbought)
```

```
##      Record Gender No_kids Education       HasCar Income PropertyValue
## 1 Record329   Male       0  Graduate           No   2727         47200
## 2 Record330 Female       0  Graduate           No   1993         43200
## 3 Record331   Male       0  Graduate           No   3580         40000
## 4 Record332   Male       0  Graduate           No   1980         37600
## 5 Record335 Female       0  Graduate           No   3561         24000
## 6 Record338   Male       0  Graduate Not Answered  69671        392000
##   Loan_Period Credit_Record Housing_type Property_Purchased
## 1         336             1      Premium                  N
## 2         336             1      Premium                  N
## 3         336             1      Premium                  N
## 4         336             1      Premium                  N
## 5         336             1      Premium                  N
## 6         156             1      Premium                  N
```

```
nrow(notbought)
```

```
## [1] 177
```

#Arrange the data frame notbought by Income and print the first five rows.

```
by_incomenb <- notbought%>%arrange(Income)
head(by_incomenb)
```

```
##      Record Gender No_kids    Education HasCar Income PropertyValue Loan_Period
## 1 Record370   Male       0     Graduate     No    165        108000         336
## 2 Record468   Male       1     Graduate    Yes   1100         88000         336
## 3 Record349   Male       2     Graduate     No   1429         13600          96
## 4 Record462 Female       2     Graduate     No   1516        133600         336
## 5 Record479   Male       0 Not Graduate     No   1587         28000         336
## 6 Record447 Female       0     Graduate     No   1650         82400         336
##   Credit_Record Housing_type Property_Purchased
## 1             1   Affordable                  N
## 2             1      Premium                  N
## 3             1      Premium                  N
## 4             1      Premium                  N
## 5             1      Premium                  N
## 6             0    Mid Range                  N
```

#Arrange the data frame notbought by Gender and print the first five rows.

```
by_gendernb <- notbought%>%arrange(Gender)
head(by_gendernb)
```

```
##      Record Gender No_kids Education       HasCar Income PropertyValue
## 1 Record330 Female       0  Graduate           No   1993         43200
## 2 Record335 Female       0  Graduate           No   3561         24000
## 3 Record354 Female       0  Graduate           No   5500        120800
## 4 Record361 Female       0  Graduate           No  11000        180000
## 5 Record362 Female       0  Graduate          Yes   8186        155200
## 6 Record371 Female       0  Graduate Not Answered   3760        108000
##   Loan_Period Credit_Record Housing_type Property_Purchased
## 1         336             1      Premium                  N
## 2         336             1      Premium                  N
## 3         456             1   Affordable                  N
## 4         336             1   Affordable                  N
## 5         336             1   Affordable                  N
## 6         336             1   Affordable                  N
```

#Arrange the data frame notbought so that Gender and Education is grouped and print the first five rows.

```
by_educationnb <- bought_property%>%arrange(Gender,Education)
head(by_educationnb)
```

```
##     Record Gender No_kids Education HasCar Income PropertyValue Loan_Period
## 1  Record1 Female       0  Graduate     No    710         90400         456
## 2 Record41 Female       0  Graduate     No   4139        108000         336
## 3 Record42 Female       0  Graduate     No   5500        105600         336
## 4 Record53 Female       0  Graduate     No   7920         96000         336
## 5 Record79 Female       0  Graduate     No   3190         56800         336
## 6 Record81 Female       0  Graduate     No   2758         44800         336
##   Credit_Record Housing_type Property_Purchased
## 1             1   Affordable                  Y
## 2             1   Affordable                  Y
## 3             1   Affordable                  Y
## 4             1   Affordable                  Y
## 5             1   Affordable                  Y
## 6             1   Affordable                  Y
```

#Reverse the order of arranging -Arrange the housing data according to Gender and decreasing Income. Print the first five rows.

```
by_reverse <- notbought%>%arrange(Gender,desc(Income))
head(by_reverse)
```

```
##      Record Gender No_kids Education HasCar Income PropertyValue Loan_Period
## 1 Record334 Female       1  Graduate     No  14589         32000         336
## 2 Record485 Female       1  Graduate    Yes  12650        228800         336
## 3 Record361 Female       0  Graduate     No  11000        180000         336
## 4 Record417 Female       0  Graduate     No  11000        171200         336
## 5 Record362 Female       0  Graduate    Yes   8186        155200         336
## 6 Record404 Female       0  Graduate    Yes   6050         84000         336
##   Credit_Record Housing_type Property_Purchased
## 1             1      Premium                  N
## 2             0      Premium                  N
## 3             1   Affordable                  N
## 4             1    Mid Range                  N
## 5             1   Affordable                  N
## 6             0   Affordable                  N
```

5.Summarise function: •Print out a summary with variables min_income and max_income.•Generate summary statistics about Income column of housing dataframe. The summary should print minimum, maximum, average, standard deviation, and IQR of the variable.•Generate summary about PropertyValue column of housing. The output should print minimum, maximum, average, standard deviation, and IQR of the variable.•Generate summary about Loan_Periodcolumn of housing. The output should print minimum, maximum, average, standard deviation, and IQR of the variable.

Print out a summary with variables min_income and max_income.

```
max_income <- housing_data%>%summarise(max(Income))
max_income
```

```
##   max(Income)
## 1       89100
```

```
min_income <- housing_data%>%summarise(min(Income))
min_income
```

```
##   min(Income)
## 1         165
```

#Generate summary statistics about Income column of housing dataframe. The summary should print minimum, maximum, average, standard deviation, and IQR of the variable.

```
summary(housing_data$Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     165    3185    4217    5953    6400   89100 
```

#Generate summary about PropertyValue column of housing. The output should print minimum, maximum, average, standard deviation, and IQR of the variable

```
summary(housing_data$Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     165    3185    4217    5953    6400   89100 
```

#Generate summary about Loan_Periodcolumn of housing. The output should print minimum, maximum, average, standard deviation, and IQR of the variable.

```
summary(housing_data$Loan_Period)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    12.0   336.0   336.0   317.9   336.0   456.0 
```

6.the pipe operator of dplyr: reproduce the below steps using dplyr and pipe operator•Start with the housing data set and then•Add the new variable var1 which calculates the ratio of property value to total income•Pick all-of the rows whose var1 value exceeds 50, and then•Summarize the data set with a value named avg. that is the mean value of var1.•Finally report the output of the above steps.

```
housing_data%>%mutate(var1 = PropertyValue/Income)%>%filter(var1 > 50)%>%summarise(mean(var1))
```

```
##   mean(var1)
## 1   112.4228
```

7.using group_by function of dplyr: reproduce the below steps •Start with the housing data set and then Use group_by() to group housing by Education.•summarise() the grouped df with two summary variables: avg_income, the average of Income, and avg_Value, the average value of purchased property. •Finally, order the summary from low to high by these two summarized variables•Finally report the output of the above steps.

```
housing_data%>%group_by(Education)%>%summarise(avg_income = mean(Income), avg_value = mean(PropertyValue))%>
%arrange(desc(avg_income),desc(avg_value))
```

```
## # A tibble: 2 x 3
##   Education    avg_income avg_value
##   <fct>             <dbl>     <dbl>
## 1 Graduate           6391.   121232.
## 2 Not Graduate       4242     93880.
```