PES University, Bengaluru
UE18CS312 - Data Analytics

Session: Aug – Dec 2020
Weeks 1-2 – Worksheet 1(a) (for Unit 1)

Dataset : BKB.csv
Source : Business Analytics, U. Dinesh Kumar
Libraries : *ggplot2, dplyr, plyr, corrplot ,e1071*
R Basics : The R Project for Statistical Computing: R
Relevant Courses/Content : Udemy
CRAN
R Programming for Data Science Roger D Peng
Compiled by: Ms. Bharani Ujjaini Kempaiah, Mr. Ruben John and Ms. Bhavya Charan
VII CSE, PES University RR Campus

Questions to explore

Getting started

1. **Read the BKB.csv dataset**
2. **Find a basic summary of the data**

Descriptive Statistics

3. **Are there any outliers in these variables? Plot a box and whisker plot to find out.**
4. **Visualize the Loan Amount attribute (Histogram is suggested, why?)**
    - Try changing the bin width of the histogram
    - You can see that since the bin width influences the nature of the distribution of a histogram, in order to find the modality of the distribution, plot a **density plot**.
    - Which other visualization is suitable for the Loan Amount Variable?

Confidence Interval and Hypothesis Testing

5. **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?**

Visualizations

6. **Visualize the distribution of Accomodation Type attribute (PieChart is suggested)**
7. **Visualize the Gender attribute (Bar Graph is suggested)**

8. **Find variation of Monthly Salaries with respect to EMI amount (Scatter Plot is suggested)**
    - Is there a significant trend in the plot? Does lower income imply lower loan amount requested?
    - Try plotting scatter plot matrices where you can visualize multiple variables at once

Summary Statistics and Grouping Conditions

9. **Find general descriptive statistics(mean, median, mode, range, standard deviation etc.) of the Monthly Salary attribute**
    - R does not have an inbuilt function for Mode. Try writing one by yourself.

10. **Find the mean monthly salary for females**
    - Try finding the **median** of Monthly Salary for Males

11. **Find the mean monthly salaries, grouped by the Gender attribute. Explore the dplyr package**
    - Try to get median and range of Monthly salary grouped by the Gender attribute

12. **Find the Skewness and kurtosis for the Monthly Salary attribute**
    - Is the attribute left skewed?
    - What about its kurtosis? Platykurtic? Leptokurtic?

Correlation and Data Reduction

13. **Find the value of correlation between Loan amount and Down payment**

14. **Explore the corrplot package to plot a correlogram between the various attributes**

15. **Perform PCA on the data**

16. **Try plotting a visualization depicting PCA (Explore ggbiplot)**

*****