**UNIT 3: DATA SUMMARY AND VISUALIZATION**

**Slide 12 - 17**

**Work 1:** Use “airquality” data of R and locate median of “Temp” variable graphically

Hint:

1. Divide the “Temp” variable into different class intervals using a statistical rule and get number of frequencies in each class interval
2. Get less than frequency data for less than ogive
3. Get more than frequency data for more than ogive
4. Plot less than and more than ogives in a single plot
5. Intersection of less than and more than ogive in the x-axis is the median
6. Check this value with median code of R

**Work 2:** Use “airquality” data of R and locate mode of “Temp” variable graphically

Hint:

1. Get histogram of “Temp” variable
2. Draw a diagonal line from en edge of the largest bar to the tip of the opposite adjacent bar
3. Draw another diagonal line from other edge of the largest bar to the tip of of the opposite adjacent bar
4. Intersection of the two diagonal lines in the x-axis in the mode
5. Check this value with mode code of R

**Work 3:** Use “SNA\_School.csv” data and perform social network analysis of first and second variables

Hint:

1. Import and create a data frame “s” with first and second column of the data
2. Save it as network graph data object “net” with directed = T argument
3. Check number of vertices, edges, degree of “net” and interpret the carefully
4. Get histogram of net degree and interpret it carefully
5. Get network diagram of “net” and interpret it carefully
6. Get network diagram of “net” with kamada.kawai layout and interpret it carefully
7. Get hubs using hubs score and interpret it carefully
8. Get authority using authority score and interpret it carefully
9. Get community using the a special network diagram parameter and interpret it carefully

**UNIT 4: SUPERVISED LEARNING**

**Slide 18 - 25**

**Part I:**

Do the following in RStudio using “airquality” dataset with R script to knit PDF output:

a) Get mean and standard deviation of Wind variable by Month variable using the appropriate “apply” family of function, show both the results in a single table and interpret them carefully

b) Perform goodness-of-fit test on Wind variable by Month variable to check if it follows normal distribution or not

c) Perform **goodness-of-fit** test on Wind variable by Month variable to check if the variances of mpg are equal or not on am variable categories

d) Discuss which **one-way ANOVA** must be used to compare “Wind” variable by “Month” variable categories based on the results obtained above

e) Fit the best one-way ANOVA for this data now and interpret the results carefully

f) Fit the most-appropriate post-hoc test if the ANOVA is statistically significant and interpret the result carefully

**Part II:**

Do the following in RStudio using “USArrests” dataset with R script to knit PDF output:

1. Create an “crime” dataset containing all the variables of USArrests
2. Create correlation matrix plot of the crime data and interpret each scatterplot carefully
3. Randomly split the crime dataset into training and testing data with 70% and 30% cases
4. Fit a **multiple linear regression** on training data with Murder as dependent variable and all other variables as independent variables and interpret the results carefully using R-squared, RMSE, Regression ANOVA and Regression Coefficients (BLUE?)
5. Check multicollinearity and finalize this model with the appropriate VIF cut-off value
6. Perform residual analysis of this model i.e. LINE tests using suggestive graphs and confirmatory tests and interpret the results carefully
7. Predict the Murder in the testing dataset using the fitted model
8. Report R-square and RMSE of predicted model and interpret them carefully
9. Report R-square and RMSE of predicted model using other **cross-validation methods** i.e. LOOCV, k-fold and repeated k-fold
10. Which predicted model is the best model? Why?
11. Fit **KNN** and **ANN**-**MLP** regression on the training data and predict on the testing data.
12. Compare R-square and RMSE of KNN and ANN-MLP models with linear regression
13. Which regression model is best? Why?

**Part III**

Do the following in R Studio using “mtcars” dataset with R script to knit PDF output:

**a)** Divide the mtcars data into train and test datasets with 80:20 random splits

**b)** Fit a supervised **logistic regression** model classification model on train data with “am” as dependent variable and all other variables as independent variable

**c)**  Check multicollinearity of this model and finalize it using appropriate VIF cut-off value for logistic regression

**d)** Get the confusion matrix, sensitivity, specificity of the fitted model and interpret them carefully

**e)** Predict the transmission variable in the test data and interpret the result carefully

**f)** Get the confusion matrix, sensitivity, specificity of the predicted model and interpret them carefully

**g)** Fit a supervised **naïve bayes, support vector machine, decision tree, bagging, random forest, tuned random forest model** and **boosting** classification models on training data with “am” as dependent variable and all other variables as independent variables

h) Which supervised classification model is the best predictive model? Why?

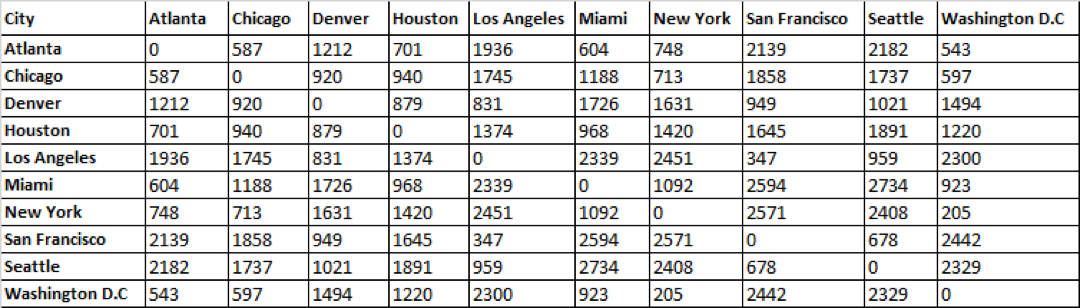
UNIT 5: UNSUPERVISED LEARNING

Slide 26 - 29

**Part I:** Use NCI60 data of ISLR2 package and page 540 of ISLR2 book to do as follows in R Studio to knit PDF output:

1. Define nci labels (NCI$labs) as nci.labs and nci data (NCI$data) and nic.data
2. Check dimension of nci.data object and interpret it carefully
3. Check first four cancer types using nci.labs object
4. Fit **principal component analysis (PCA)** on nci.data with scale = TRUE argument as pr.out object
5. Create a plot showing first three PCA components with three different colors
6. Get summary of pr.out object and interpret it carefully
7. Plot pro.out object and interpret it carefully
8. Create custom scatterplots with principal components in x-axis and proportion variance explained (PVE) in y-axis for the first plot and cumulative PVE in the y-axis for the second plot and interpret them carefully
9. Perform PCA with varimax rotation and compare it with the PCA result obtained above
10. Write summary of the results and conclusion based on your findings

**Part II:** Use the distance between 10 US cities provided below in R Studio to knot PDF output:



1. Get dissimilarity distance as city.dissimilarity object
2. Fit a classical multidimensional model using the city.dissimilarity object
3. Get the summary of the model and interpret it carefully
4. Get the bi-plot of the model and interpret it carefully

**Part III:** Use NCI60 data of ISLR2 package and page 543 of ISLR2 book to do as follows in R Studio to knit PDF output:

1. Scale the nci.data as sd.data object
2. Fit **hierarchical cluster analysis** on the sd.data using complete, average and single linkage methods, show the results with dendrogram and interpret them carefully
3. Find the best number for clusters using “cutree” function with best distance value
4. Use your roll number as set.seed and perform **k-means clustering** on sd.data with the best number of clusters/distance value with nstart=20
5. Get summary of the k-means clustering and interpret them carefully
6. Plot this k-means results using base r plot and cluster package and interpret them carefully

**Part IV:** Use “Groceries” data available in the “datasets” package to do as follows in R Studio to knit PDF output

1. Load “arules” and “arulesViz” libraries
2. Load “Groceries” data, check its structure and interpret it carefully
3. Get Frequent Item frequencies using itemFrequencyPlot function and interpret it carefully
4. Set a **priori rule** with support = 0.001 and confidence = 0.8 and interpret the output carefully
5. Show the top five rules using inspect and round the results to two digits
6. Sort the rule by confidence in decreasing order
7. Use “whole milk” as target item and show the items in “lhs” with decreasing order of confidence and show the top five rules
8. Use “whole milk” as target item and show the items in “rhs” with decreasing order of confidence and show the top five rules
9. Write summary and conclusion based on your findings above

PRESENTATION SLIDE

Q no 1

Roll Number: 1, 5, 9, 13, 17, 21, 25, 29, 33

Generate 200 random data with one continuous dependent variable and five independent variables, random seed = your roll number.

Then fit **bivariate and multivariate linear regression** model,

check VIF, BLUE and LINE and predict for the test data (30% sample)

compare it with **KNN regression, ANN regression and Support Vector Regression** using R-square and RMSE of test data.

find the **best supervised regression model** for this data.

Q no 2

Roll Number: 2, 6, 10, 14, 18, 22, 26, 30, 34

Generate 200 random data with one binary/dichotomous dependent variable and two continuous independent variables and three categorical independent variables, random seed = your roll number.

Then fit **bivariate and multivariate logistic regression** model,

check VIF and BLUE and predict it for test data (20% sample) and

also compare it with **Naive Bayes, SVM, Decision Tree, Bagging, Random Forest, Tuned Random Forest** and **Boosting** using confusion matrix of test data and

find the **best supervised classification model** for this data.

Q no 3

Roll Number: 3, 7, 11, 15, 19, 23, 27, 31, 35

Generate 100 random data with ten continuous independent variables, random seed = your roll number.

Then fit PCA, PCA with Varimax rotation, MDS, MDS with Shannon's stress minimization with Bi-plots and interpret them carefully.

Compare final **PCA** with final **MDS** model graphically.

Finally, find the **best unsupervised dimension reduction model** for this data.

Q no 4  
Roll Number: 4, 8, 12, 16, 20, 24, 28, 32

Generate 100 random data with seven continuous independent variables where first independent variable must be the identification variable with name (text), random seed = your roll number.

Then fit **Hierarchical Cluster Analysis (HCA)** with four methods

decide how many clusters (k) must be used for the k-means clustering for each method.

Fit **k-means clustering** for all the methods using the pre-identified k values for each method.

Finally, find the **best unsupervised clustering model** for this data and

show it using **base r graph** and **cluster** package.