Program	Post Graduate Diploma in Data Science
Semester	1
Subject Code and Name	1618002 Statistics and Exploratory Data Analysis
Credit	5

Objectives

- To impart basic managerial skills for collection & analysis of statistical data.
- To learn about graphical and modelling techniques for exploring data, with an emphasis on visualization, interpretation, and clear communication of findings.

Unit No.	Topic(s)	No. of Hours
1.	**Introduction to Statistical Methods** Statistics & Managerial Decisions, Statistical Data, Operation Research Techniques	4
2.	**Data Collection And Analysis** Collection and presentation of data in terms of tables, graphs, raw data, frequency distributions, histogram etc. Cumulative frequency curve, Measures of central tendency and location, Partition values, Comparison of various measures of central tendencies, Measures of dispersion, skewness & kurtosis, comparison of various measures of dispersion	8
3.	**Probability Distribution & Statistics** Introduction of Probability, sample, space & events, Basic rules of probability, permutation & combinations, conditional probability, Bayes' theorem, distributions: Binomial, Poisson, Exponential and Normal distribution with their properties and application. Random variables – discrete and continuous probability distribution functions	8
4.	**Correlation And Regression Analysis** Curve fitting, correlation and regression analysis, Autocorrelation, Multiple regression, statistical Inference & estimation applied to Industrial problems	6
5.	**Understanding Data for Exploratory Analysis** Exploratory data analysis and data visualization, Perception, Continuous variables, Discrete variables, Dependency relationships, Multivariate categorical variables, Temporal data, Spatial data Data Science Pipeline: Collect, Import, Clean, Transform, Visualize, Model, Communicate	10
6.	**Statistical Tests and Testing of Hypothesis** Elementary theory and practice of sampling, standard error or means and variance, tests of significance, T test, F test, Z test and chi-square test along with their applications, Goodness of fit, testing of hypotheses and decision making, analysis of variance (ANOVA)	4

Reference Books

1. Quantitative Techniques for Managerial Decision
   by U. K. Srivastava, G. V. Shenoy and S. C. Sharma
   New Age International Publishers
2. Probability & Statistics for Engineers
   by Rao
   SCITECH
3. Statistics for Management
   by Lewis
   Pearson
4. Graphical Data Analysis with R
   by Unwin, Antony
   CRC Press, 2015 ISBN 978-1498715232
5. Interactive Data Visualization for the Web
   by Scott Murray
   O'REILLY, Second edition, ISBN-13: 978-1491921289

Outcomes

After completion of this subject, students would be able to:

- Identify objectives of statistical analysis.
- Apply methods of data collection & analysis.
- Use correlation and regression analysis.
- Visualize the data to get the insights from it further for decision making process.

Suggested list of Practical (at least 10 practical are to be performed by students. These practical should cover majority of all topics of syllabus.)
This is the suggested list of practical but it may not be limited only to this list.

1. Download a data-set from uci repository/kaggle which has data of mix type (i.e. string, float, date,). Perform following operations in R program.
   - (i) Import the data, create the data frame
   - (ii) View the data
   - (iii) Print first few and last few records
   - (iv) Create a data frame which holds subset of original data
   - (v) Create a vector which holds values of one of the columns

2. (i) Create a vector having following marks: 56, 89, 76, 54, 79, 90, 75, 48.
   (ii) Create a vector having following students: Karan, Rakesh, Hiren, Rudra, Himesh, Shantanu, Rohan, Sidhdhesh.
   (iii) Create a data frame from above two vectors.
   (iv) Update the marks of Rakesh by more 7 marks.
   (v) Retrieve the students who secured more than 85% of marks.
   (vi) Sort the student list by descending order of their marks.

3. Import the GDP dataset from kaggle and compute the difference in GDP between 2007 and 2017 for each country. Also create a subset of countries that saw an increase of over one trillion dollars.

4. For the GDP dataset, compute the measures of central tendency, mean, median, range and quantile for year 2017.

5. Import Winter_olympics dataset from any of the public dataset repository.
   (i) Sort data by total medals and country and save it in new data frame.
   (ii) Check mean and median of number of gold, silver, bronze and total medals.
   (iii) Find that which region won the highest mean total medals?
   (iv) What is the maximum number of medals won? Which country won it?
   (v) Find correlations between total medals and number of gold and bronze.
   (vi) Find correlation between rank and total medals.

6. Import Movies dataset from kaggle/uci public dataset repository.
   (i) Make scatter plot of tickets sold versus gross.
   (ii) Redo the scatter plot, adjusting scales, divide by 1000, 100,000 and 1,000,000
   (iii) Find the correlation between tickets sold and sales? Is it expected?
   (iv) Make scatter plot with millions scale and also add a regression line with label to all axes and title to plot.
   (v) Make boxplot
   (vi) Make individual histogram of type of films, gross sales and ticket sales.

7. For the GDP and Life Expectancy datasets from kaggle, produce following plots that describe the GDP and Life Expectancy during the year 2016.
   (i) Create a scatter plot of GDP to Life Expectancy
   (ii) Create a histogram of GDP
   (iii) Create a box and whisper plot of Life Expectancy

8. Import Summer_Winter_olympics dataset from any of the public dataset repository.
   (i) Look at the column names and change names to more meaningful names.
   (ii) Make two histograms on one page: summer games (total), winter games (total).
   (iii) Make two histograms on one page: total summer, total winter medals won
   (iv) Check that is there a correlation between number of medals won in winter and summer?
   (v) Check that is there correlation between number of games each country competes in winter and summer.
   (vi) Make 6 histograms on one page showing the distribution of each of the types of medals by seasons.

9. For the GDP, Life Expectancy and Employment datasets from kaggle, perform following tasks.
   (i) Merge the columns for the year 2016 for 3 datasets into a new data frame
   (ii) Rename the columns to “country”, “gdp”, “life_expectancy” and “employment”
   (iii) Create a frequency table for each variable
   (iv) Draw histograms for each variable

10. Import any real life data set from any public data set repository. Create a shiny app that allows the user to interactively display a box plot based on selection of value of other column.

****