

SYLLABUS

CS3352 - FOUNDATIONS OF DATA SCIENCE

UNIT - I: INTRODUCTION

9

Data Science: Benefits and uses – facets of data - Data Science Process: Overview
– Defining research goals – Retrieving data – Data preparation - Exploratory Data analysis – build the model – presenting findings and building applications - Data Mining - Data Warehousing – Basic Statistical descriptions of Data.

UNIT - II: DESCRIBING DATA

9

Types of Data - Types of Variables - Describing Data with Tables and Graphs – Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores

UNIT - III: DESCRIBING RELATIONSHIPS

9

Correlation – Scatter plots – correlation coefficient for quantitative data – computational formula for correlation coefficient – Regression – regression line –least squares regression line – Standard error of estimate – interpretation of r^2 – multiple regression equations – regression towards the mean.

UNIT - IV: PYTHON LIBRARIES FOR DATA WRANGLING

9

Basics of Numpy arrays – aggregations – computations on arrays – comparisons, masks, boolean logic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets – aggregation and grouping – pivot tables.

UNIT - V: DATA VISUALIZATION

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting – Geographic Data with Basemap – Visuali with Seaborn.

9

CONTENTS

UNIT – I: INTRODUCTION

1.1	Data Science	1.1
1.2	Benefits and Uses.....	1.1
1.3	Facets of Data	1.4
1.4	Data Science Process: Overview.....	1.8
1.5	Defining Research Goals	1.9
1.6	Retrieving Data.....	1.11
1.6.1	Start with data stored within the company.....	1.11
1.6.2	Don't be afraid to shop around.....	1.12
1.6.3	Do data quality checks now to prevent problems later	1.13
1.7	Data Preparation	1.13
1.7.1	Data Entry Errors	1.16
1.7.2	Redundant Whitespace.....	1.17
1.7.3	Outliers.....	1.17
1.7.4	Dealing with Missing Values	1.18
1.7.7	Deviations from a Code Book	1.19
1.7.5	Different Units of Measurement.....	1.20
1.7.6	Different levels of Aggregation.....	1.20
1.8	The Different Ways of Combining Data	1.21
1.8.1	Using views to simulate Data joins and Appends.....	1.23
1.8.2	Enriching Aggregated Measures	1.24
1.9	Transforming Data.....	1.25
1.9.1	Reducing the Number of Variables	1.26
1.9.2	Turning Variables into Dummies	1.26
1.10	Exploratory Data Analysis.....	1.27
1.10.1	Brushing and linking.....	1.28
1.10.2	Histogram and Boxplot	1.29
1.11	Build The Models	1.30
1.12	Presenting Findings and Building Applications on Top of Them.....	1.35

1.13	Data Mining.....	1.36
1.13.1	Data Mining Application.....	1.37
1.13.2	Data Mining Techniques.....	1.38
1.14	Data Warehousing.....	1.39
1.14.1	Types of Data Warehouse.....	1.40
1.14.2	Stages of Data Warehouse.....	1.40
1.14.3	Components of Data warehouse.....	1.41
1.14.4	Who needs Data warehouse?.....	1.41
1.14.5	Applications of Datawarehouse.....	1.41
1.14.6	Advantages of Datawarehouse.....	1.42
1.14.7	Disadvantages of Datawarehouse.....	1.42
1.15	Data Warehouse Tools.....	1.42
1.16	Basic Statistical Descriptions of Data.....	1.43
1.17	Coefficient of Determination, r^2	1.44
1.17.1	Descriptive Statistics.....	1.47
1.17.2	Inferential Statistics.....	1.48
1.17.3	Populations and Samples.....	1.49
1.17.4	Random Sampling (Surveys).....	1.49
1.17.5	Random Assignment (Experiments).....	1.49
	Part B – Questions.....	1.50

UNIT – II: DESCRIBING DATA

2.1	Introduction - Data.....	2.1
2.1.1	Levels of measurement.....	2.3
2.2	Types of Variables.....	2.8
2.3	Describing Data With Tables and Graphs.....	2.10
2.3.1	Frequency distributions for ungrouped Data.....	2.10
2.3.2	Frequency Distribution for Grouped Data.....	2.11
2.3.3	Guidelines for frequency distributions.....	2.11
2.3.4	Constructing Frequency Distributions.....	2.14
2.4	Outliers.....	2.15
2.5	Relative Frequency Distributions.....	2.15

2.5.1	Constructing Relative Frequency Distributions	2.16
2.5.2	Cumulative frequency distributions	2.17
2.6	Frequency Distributions for Qualitative (Nominal) Data.....	2.18
2.6.1	Graphs.....	2.18
2.6.2	Histograms	2.19
2.6.3	Frequency Polygon.....	2.20
2.7	Stem and Leaf Displays	2.22
2.7.1	Constructing a Display	2.22
2.7.2	Interpretation.....	2.23
2.8	A Graph for Qualitative (Nominal) Data.....	2.25
2.8.1	Misleading graphs	2.26
2.8.2	Constructing graphs	2.27
2.9	Normal Distributions and Standard (Z) Scores	2.28
2.9.1	Properties of the Normal Curve.....	2.28
2.9.2	Importance of Mean and Standard Deviation.....	2.29
2.9.3	Different Normal Curves.....	2.29
2.9.4	Standard normal curve	2.31
2.9.5	Standard Normal Table	2.32
2.9.6	Finding proportions.....	2.32
2.10	Describing Data With Averages.....	2.38
2.11	Which Average?	2.41
2.11.1	If Distribution Is Not Skewed.....	2.41
2.11.2	If Distribution Is Skewed	2.42
2.11.3	Interpreting Differences between Mean and Median.....	2.43
2.12	Averages for Qualitative and Ranked Data.....	2.45
2.13	Describing Variability.....	2.45
2.13.1	Range	2.46
2.13.2	Variance	2.46
2.13.3	Standard Deviation.....	2.47
2.13.4	Details: Standard Deviation.....	2.50
2.14	Calculation of The IQR.....	2.52

UNIT – III: DESCRIBING RELATIONSHIPS

3.1	Introduction - Correlation	3.1
3.2	Scatter Plots	3.4
3.2.1	Positive, Negative, or Little or No Relationship.....	3.5
3.2.2	Strong or Weak Relationship?.....	3.5
3.2.3	Perfect Relationship	3.6
3.2.4	Curvilinear Relationship	3.7
3.3	Correlation Coefficient for Quantitative Data	3.7
3.3.1	Key Properties of r	3.8
3.3.2	Sign of r	3.8
3.3.3	Numerical Value of r	3.8
3.3.4	Interpretation of r	3.9
3.3.5	r is Independent of Units of Measurement.....	3.9
3.3.6	Range Restrictions	3.9
3.3.7	Verbal Descriptions.....	3.11
3.3.8	Correlation Not Necessarily Cause-Effect	3.11
3.3.9	Role of Experimentation	3.12
3.4	Outliers	3.17
3.5	Regression	3.19
3.6	Regression Line	3.21
3.6.1	Least Squares Regression Line.....	3.25
3.6.2	Determining the Least Squares Regression Equation.....	3.26
3.6.3	Standard Error of Estimate	3.27
3.7	Interpretation of r^2	3.28
3.7.1	Multiple Regression Equations	3.32
3.7.2	Regression Towards The Mean	3.32
	Part B – Questions	3.35

UNIT – IV: PYTHON LIBRARIES FOR DATA WRANGLING

4.1	Python Libraries for Data Wrangling.....	4.1
-----	--	-----

4.1.1	Introduction - Basics of NumPy Arrays	4.3
4.1.2	NumPy Array Attributes	4.3
4.1.3	Array Indexing: Accessing Single Elements	4.4
4.1.4	Array Slicing: Accessing Subarrays	4.5
4.1.5	Reshaping of Arrays	4.7
4.1.6	Array Concatenation and Splitting	4.8
4.2	Aggregations	4.9
4.2.1	Summing the Values in an Array	4.9
4.2.2	Minimum and Maximum	4.9
4.2.3	Multidimensional aggregates	4.10
4.3	Computations on Arrays	4.10
4.3.1	Exploring NumPy's UFuncs	4.11
4.3.2	Summing the Values in an Array	4.13
4.4	Comparisons, Masks, Boolean Logic	4.14
4.4.1	Working with Boolean Arrays	4.15
4.4.2	Boolean operators	4.16
4.4.3	Boolean Arrays as Masks	4.16
4.5	Fancy Indexing	4.17
4.5.1	Exploring Fancy Indexing	4.17
4.5.2	Combined Indexing	4.19
4.5.3	Modifying Values with Fancy Indexing	4.19
4.6	Structured Arrays	4.20
4.6.1	Creating Structured Arrays	4.22
4.6.2	RecordArrays: Structured Arrays with a Twist	4.23
4.7	Data Manipulation With Pandas	4.23
4.7.1	Installing and Using Pandas	4.23
4.7.2	Introducing Pandas Objects	4.24
4.7.3	The Pandas Series Object	4.24
4.7.4	Constructing Series objects	4.25
4.7.5	The Pandas Index Object	4.27
4.8	Data Indexing and Selection	4.28

4.8.1	Data Selection in Series.....	4.28
4.8.2	Series as one-dimensional array	4.29
4.8.3	Data Selection in DataFrame.....	4.31
4.9	Operating on Data.....	4.32
4.9.1	Ufuncs: Index Preservation	4.33
4.9.2	UFuncs: Index Alignment	4.34
4.10	Missing Data.....	4.36
4.10.1	NaN: Missing numerical data.....	4.37
4.10.2	NaN and None in Pandas.....	4.37
4.10.3	Operating on Null Values.....	4.38
4.10.4	Detecting null values	4.39
4.10.5	Dropping null values	4.39
4.10.6	Filling null values.....	4.41
4.11	Hierarchical Indexing	4.43
4.11.1	A Multiply Indexed Series	4.43
4.11.2	MultiIndex as extra dimension.....	4.44
4.11.3	Methods of MultiIndex Creation.....	4.45
4.12	Combining Datasets.....	4.45
4.12.1	Concatenation of NumPy Arrays.....	4.46
4.12.2	Simple Concatenation with pd.concat.....	4.46
4.12.3	Concatenation with joins.....	4.47
4.12.4	The append() method	4.47
4.12.5	Combining Datasets: Merge and Join.....	4.48
4.13	Aggregation and Grouping	4.50
4.13.1	Simple Aggregation in Pandas	4.51
4.13.2	GroupBy: Split, Apply, Combine.....	4.53
4.14	Pivot Tables.....	4.54
4.14.1	Motivating Pivot Tables.....	4.54
	Part B – Questions	4.57

UNIT – V: DATA VISUALIZATION

5.1	Introduction	5.1
5.1.1	Importing Matplotlib	5.1
5.2	Line Plots	5.2
5.3	Scatter Plots	5.9
5.4	Visualizing Errors	5.13
5.5	Density and Contour Plots	5.16
5.5.1	Visualizing a Three-Dimensional Function	5.16
5.6	Histograms	5.19
5.7	Legends	5.24
5.7.1	Choosing Elements for the Legend	5.25
5.7.2	Legend for Size of Points	5.26
5.8	Colors	5.27
5.8.1	Customizing Colorbars	5.28
5.8.2	Choosing the colormap	5.29
5.8.3	Color limits and extensions	5.30
5.8.4	Discrete colorbars	5.31
5.9	Subplots	5.32
5.9.1	plt.subplot: Simple Grids of Subplots	5.33
5.10	Text and Annotation	5.35
5.11	Customization	5.37
5.11.1	Major and Minor Ticks	5.37
5.11.2	Hiding Ticks or Labels	5.38
5.11.3	Plot Customization by Hand	5.39
5.12	Three Dimensional Plotting	5.40
5.13	Geographic Data With Basemap	5.44
5.14	Visualization With Seaborn	5.51
5.14.1	Seaborn Versus Matplotlib	5.51
5.14.2	Exploring Seaborn Plots	5.53
	Part B – Questions	5.55
	Two Marks Questions and Answers	1-34
	Solved Question Paper	