**UNIT 1**

**Business Analytics:**

**Definition**: -“Process of transforming data into actions through analysis and insights in the context of organisational decision making and problem solving” -It is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain an improved insight about their business operations and make better, fact-based decisions -Supported by various tools such as Microsoft excel, and other software packages

**Importance of Analytics:** -Data, facts and analysis aid decision making, and that the decisions made on them are better than those made through gut instinct -Decision making today is even more complicated, due to overwhelming data and information -There is a strong relationship of use of analytics and profitability and revenye

**Evolution of Business Analytics:** -Modern evolution of analytics began with the introduction of computers, as they provided the ability to store and analyze data easily.

Three major components of business analytics:

1. **Descriptive Analysis** -Most commonly used and most well understood type of analytics (WANT TO KNOW -Use data to understand past and present performance to make important ABOUT PAST) decisions -Summarizes data into meaningful charts and reports

2. **Predictive Analysis** -Analyzes past performance in an effort to predict the future by examining (WANT TO KNOW historical data, detecting patterns or relationships in these data ABOUT FUTURE) -Techniques include: regression and forecasting

3. **Prescriptive Analysis** -Uses optimization to identify the best alternative to minimize or maximize (MAKING DECISIONS- some objective OPTIMIZATION) -Addresses questions such as: •How much should we produce to maximize profit? •What is the best way of shipping goods from our factory to minimize costs?

**What is Statistics?**

Statistics definition: -"Statistics relates to the collection, analysis, interpretation, and presentation of data" -Statistical methods are used to:

•Summarize a collection of data

•Draw inferences about an entire population

•Make predictions or forecasts -Statistics is also the study of variation in data

-**Descriptive VS. Inferential statistics:**

1. Descriptive statistics: -Are tabular, graphical, and numerical measures used to summarize data 2. Inferential statistics: -The process of using data obtained from a sample to make estimates and test claims about the characteristics of a population

Variables: -Characteristics of items or individuals -EG. Gender, field of study, money in wallet, time spent in shower each day -It is essential that all variables have an operational definition: which is defines how a variable is to be measured, otherwise confusion can occur.

Data: -Observed characteristics of items of individuals.

Populations: -A collection of all members of a group being investigated -Two factors need to be specified when defining a population:

•1. The entity (EG. People or motor vehicles)

•2. The boundary

Sample: -The portion of the population selected for analysis -EG. Ten full time students selected for a focus group

Parameter: -A numerical measure of some population characteristic

-EG. The average amount spent by all customers at the local shopping centre last weekend

Statistic: -A numerical measure that describes a characteristic of a sample -EG. The average amount spent by the 30 customers completing the market research survey

**Data sources:**

Four important sources of data: -Data distributed by an organisation or an individual -A designed experience -A survey -An observational study (such as a focus group)

**Primary and Secondary sources:**

Primary sources: -When the data collector is the one using the data for analysis -EG. Internal company records, business transactions, customer market surveys Secondary sources: -When another organisation or individual has collected the data that is used for analysis by an organisation or individual -EG. Government and commercial sources, online research

**Types of Data:**

\*BIG DATA\* (Data deluge): -Many companies have massive amounts of data at their disposal

-This data deluge is a result of:

•Automatic data collection

•Electronic instrumentation

•Online transactional processing -

There is growing recognition of the untapped value in these data bases -

Data is produced in great volumes, in a variety of forms, and is produced very quickly=

BIG DATA

1. Categorical data (Qualitative data): -

Labels or names used to identify attributes of each entity -

Can be recorded in either numeric or nonnumeric formats -

EG. 'Yes or no', 'male or female' answers

-Usually counted or expressed as a portion or a percentage

2. Numerical data (Quantitative data): -Take numbers as their observed responses -Numerical data can be converted to categorical data. EG Salary can be converted into low/medium/high. However you cannot convert categorical data back to numerical data -There are two types of numerical data:

Discrete:

-If measuring how many (Whole numbers)

Continuous:

-If measuring how much (Decimal places)

Scales of Measurement:

Categorical Measurements

**Nominal**:

-A classification of categorical data that implies no ranking -

EG. Favorite soft drink, gender

**Ordinal**: -

Scale of measurement where values are assigned by ranking

-EG. Rating customers service as 'very good, good, average, or poor' Numerical Measurements

**Interval**:

-A ranking of numerical data where differences are meaningful but there is no true zero point -

EG. Shoe sizes 9, 9.5, 10

**Ratio**: -A ranking of numerical data where differences between measurements involve a true zero point -EG. Length, weight, age, salary measurements

**Z score** -"Provide a relative measure of the distance an observation is from the mean (in terms of standard deviations)"

z =X — X/ S D

-As a general rule a Z score above +3 or below -3 is considered an outlier

-EG. A Z score of 2 means that a value is 2 SDs away from the mean

**The Chebyshev Rule** (for any data set): -At least 75% of the data values must be within Z=2 Standard deviations of the mean -At least 89% of the data values must be within Z=3 Standard deviations of the mean -At least 94% of the data values must be within Z=4 Standard deviations of the mean

**The Empirical Rule** (for a data set that is bell-shaped): -Approx 68% of the data values lie within Z=1 Standard deviations of the mean -Approx 95% of the data values lie within Z=2 Standard deviations of the mean -Approx 99.7% of the data values lie within Z=3 Standard deviations of the mean (ONLY WORKS FOR SYMMETRICAL DATA)

4. Measures of Variability:

WORST MEASURE OF VARIABILITY (starting at range) BEST MEASURE OF VARIABILITY (through to coefficient of variation)

**1. Distance Measures:**

Range: -"Difference between largest and smallest data values" =Max - Min Interquartile Range (IQR): -"Difference between the third quartile and the first quartile" =Q3 - 01 -It is the range for the middle 50% of the data

2. Average Variation: -Measure the average scatter around the mean. That is how larger values fluctuate above it and how smaller values are distributed below it.

Variance (S2): -Expressed in square units . \--( v - \ )'- ' s =. Standard deviation (S):

-"Estimate of the average deviation of individual values away from the mean" -SD is preferred over S2 because it maintains the original unit -S(a) = -\/.5

3. Relative Variation:

Coefficient of Variation:

-"Indicates how large the standard deviation is in relation to the mean" cov = = x too -Useful for comparing variability between data sets in different units

-EG. Relative to the mean, the package volume is more variable than the package weight

Summary: -The more spread out the data: the larger the range + IQR + SD -The more concentrated or similar the data: the smaller the range+ IQR + SD -If the value are the same: the range + IQR + SD will be zero -No measure of variation can ever be negative

**UNIT 2**

**What is Descriptive Statistics?**

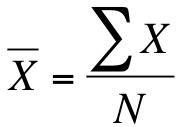
Understanding Descriptive Statistics

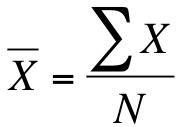
Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set. For example, the sum of the following data set is 20: (2, 3, 4, 5, 6). The mean is 4 (20/5). The mode of a data set is the value appearing most often, and the median is the figure situated in the middle of the data set. It is the figure separating the higher figures from the lower figures within a data set. However, there are less-common types of descriptive statistics that are still very important.

People use descriptive statistics to repurpose hard-to-understand quantitative insights across a large data set into bite-sized descriptions. A student's grade point average (GPA), for example, provides a good understanding of descriptive statistics. The idea of a GPA is that it takes data points from a wide range of exams, classes, and grades, and averages them together to provide a general understanding of a student's overall academic abilities. A student's personal GPA reflects his mean academic performance.

**The Mean**

The mean is what people typically refer to as "the average". It is the highest measure of central tendency, by which I mean it is available for use only with interval/ratio variables. The mean takes into account the value of every observation and thus provides the most information of any measure of central tendency. Unlike the median, however, the mean is sensitive to outliers. In other words, one extraordinarily high (or low) value in your dataset can dramatically raise (or lower) the mean. The mean, often shown as an x or a y variable with a line over it (pronounced either "x-bar" or "y-bar"), is the sum of all the scores divided by the total number of scores. In statistical notation, we would write it out as follows:





In that equation, is the mean, X represents the value of each case and N is the total number of cases. The sigma (Σ) is just telling us to add all the scores together. The fact that calculating the mean requires addition and division is the very reason it can't be used with either nominal or ordinal variables. We can't calculate a mean for race (white + white + black/3 = ?) any more than we can calculate a mean for year in school (freshman + freshman + senior/3 = ?)

**The Median**

The median is the middlemost number. In other words, it's the number that divides the distribution exactly in half such that half the cases are above the median, and half are below. It's also known as the 50th percentile, and it can be calculated for ordinal and interval/ratio variables. Conceptually, finding the median is fairly simple and entails only putting all of your observations in order from least to greatest and then finding whichever number falls in the middle. Note that finding the median requires first ordering all of the observations from least to greatest. This is why the median is not an appropriate measure of central tendency for nominal variables, as nominal variables have no inherent order. (In practice, finding the median can be a bit more involved, especially if you have a large number of observations—see your textbook for an explanation of how to find the median in such situations).

Some of you are probably already wondering, "What happens if you have an even number of cases? There won't be a middle number then, right?" That's a very astute observation, and I'm glad you asked. If your dataset has an even number of cases, the median is the average of the two middlemost numbers. For example, for the numbers 18, 14, 12, 8, 6 and 4, the median is 10 (12 + 8 = 20; 20/2 = 10).

One of the median's advantages is that it is not sensitive to outliers. An outlier is an observation that lies an abnormal distance from other values in a sample. Observations that are significantly larger or smaller than the others in a sample can impact some statistical measures in such a way as to make them highly misleading, but the median is immune to them. In other words, it doesn't matter if the biggest number is 20 or 20,000; it still only counts as one number. Consider the following:

Distribution 1: 1, 3, 5, 7, 20  
Distribution 2: 1, 3, 5, 7, 20,000

These two distributions have identical medians even though Distribution 2 has a very large outlier, which would end up skewing the mean pretty significantly, as we'll see in just a moment.

**The Mode**

The mode is the category with the greatest frequency (or percentage). It is not the frequency itself. In other words, if someone asks you for the mode of the distribution shown below, the answer would be coconut, NOT 22. It is possible to have more than one mode in a distribution. Such distributions are considered bimodal (if there are two modes) or multi-modal (if there are more than two modes). Distributions without a clear mode are said to be uniform. The mode is not particularly useful, but it is the only measure of central tendency we can use with nominal variables. You will find out why it is the only appropriate measure for nominal variables as we learn about the median and mean next.

**How to Measure Variability**

Statisticians use summary measures to describe the amount of variability or spread in a set of data. The most common measures of variability are the range, the interquartile range (IQR), variance, and standard deviation.

**The Range**

The **range** is the difference between the largest and smallest values in a set of values.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. For this set of numbers, the range would be 11 - 1 or 10.

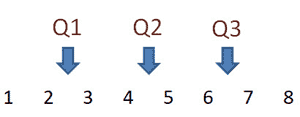
**The Interquartile Range (IQR)**

The **interquartile range** (IQR) is a measure of variability, based on dividing a data set into quartiles.

Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

* Q1 is the "middle" value in the f*irst* half of the rank-ordered data set.
* Q2 is the median value in the set.
* Q3 is the "middle" value in the *second* half of the rank-ordered data set.

The interquartile range is equal to Q3 minus Q1. For example, consider the following numbers: 1, 2, 3, 4, 5, 6, 7, 8.



Q2 is the median of the entire data set - the middle value. In this example, we have an even number of data points, so the median is equal to the average of the two middle values. Thus, Q2 = (4 + 5)/2 or Q2 = 4.5. Q1 is the middle value in the first half of the data set. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is, Q1 = (2 + 3)/2 or Q1 = 2.5. Q3 is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is, Q3 = (6 + 7)/2 or Q3 = 6.5. The interquartile range is Q3 minus Q1, so IQR = 6.5 - 2.5 = 4.

Notice that this process divided the data set into four parts of equal size. The first part consists of 1 and 2; the second part, 3 and 4; the third part, 5 and 6; and the fourth part, 7 and 8.

**The Variance**

In a population, **variance** is the average squared deviation from the population mean, as defined by the following formula:

σ2= Σ ( Xi - μ )2 / N

where σ2 is the population variance, μ is the population mean, Xi is the *i*th element from the population, and N is the number of elements in the population.

Observations from a simple random sample can be used to estimate the variance of a population. For this purpose, sample variance is defined by slightly different formula, and uses a slightly different notation:

*s*2 = Σ ( xi - x )2 / ( n - 1 )

where *s*2 is the sample variance, x is the sample mean, xi is the *i*th element from the sample, and n is the number of elements in the sample. Using this formula, the sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you need to estimate an unknown population variance, based on data from a simple random sample, this is the formula to use.

**The Standard Deviation**

The **standard deviation** is the square root of the variance. Thus, the standard deviation of a population is:

σ = sqrt [ σ ] = sqrt [ Σ ( Xi - μ )2 / N ]

where σ is the population standard deviation, μ is the population mean, Xi is the *i*th element from the population, and N is the number of elements in the population.

Statisticians often use simple random samples to estimate the standard deviation of a population, based on sample data. Given a simple random sample, the best estimate of the standard deviation of a population is:

*s*= sqrt [ s2 ] = sqrt [ Σ ( xi – x )2 / ( n - 1 ) ]

where *s* is the sample standard deviation, x is the sample mean, xi is the *i*th element from the sample, and n is the number of elements in the sample.

**Effect of Changing Units**

Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of variability are affected when we change units.

* If you add a constant to every value, the distance between values does not change. As a result, all of the measures of variability (range, interquartile range, standard deviation, and variance) remain the same.
* On the other hand, suppose you multiply every value by a constant. This has the effect of multiplying the range, interquartile range (IQR), and standard deviation by that constant. It has an even greater effect on the variance. It multiplies the variance by the square of the constant.

**Unit 3**

**Linear Regression Analysis**

Linear Regression Analysis or popularly called as, Linear analysis is a statistical technique to measure the effect and cause relationship between two variable represented by X and Y in Linear Regression Model;

* Dependent variable is Y
* Independent variable is X

Dependent variable (Y) is the one which is predicted under linear analysis and also called as criterion variable whereas

Independent variable (X) is used for predicting the x which is called as predictor variable.

If you have gone through the last chapter, there are two kinds of regression analysis, namely;

* Univariate
* Multivariate

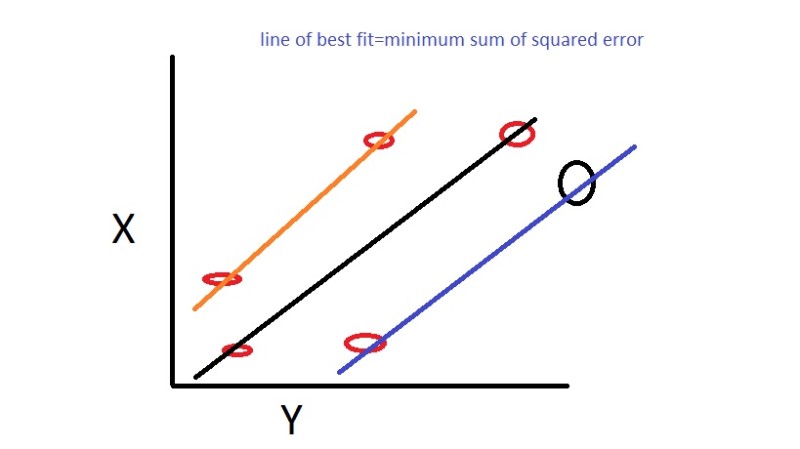
when 1 independent variable is used, analysis is called simple linear regression which comes under univariate analysis whereas when multiple variables are used in the study, it is called multiple linear analysis which falls under multivariate regression analysis.

**Line of Best Fit in Linear Analysis**

In linear analysis, first a plot is charted using the variable’s value and a line is plotted. This line is called line of best fit. It is used to join different values on the chart. You may have a situation where multiple lines could be drawn, in that case you will chose the one with minimum sum of squared errors.

You can use the graph shown here for better understanding the line of best fit. Other than line of best fit, you also need to know about;

Residual Value- It is the discrepancy between the actual and the predicted value.



Usage of R Square in Linear Regression

R2 (R Square) is based on line of best fit and is used to measure the strength of the relationship between variables in question.

* R2 is used in following;
* R2=0; No Linear Relationship
* R2=-1; Negative Linear Relationship
* R2=+1; Positive Linear Relationship

Let’s use a Linear Regression Analysis but before that let’s see how linear analysis is performed and rules to analyze it;

1. Model p-value should be less than 0.05 which have earlier seen p test and anova related chapters.
2. R Square should be greater than 0.06.
3. Adjusted R Square should be closer to R square and there should not be much difference between them.

**Data Mining**

Definition Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data.

Data Mining is defined as extracting information from huge sets of data In other words, we can co that data mining is the procedure of mining knowledge from data.

The information or knowledge extracted so can he used for any of the following applications:

1. Market Analysis

2. Fraud Detection

3. Customer Retention

4. Production Control

5. Science Exploration

**Data Mining Applications** Data mining is highly useful in the following domains:

1. Market Analysis and Management

2. Corporate Analysis and Risk Management

3. Fraud Detection

**Data Mining Techniques**

1. **Classification**: This analysis is used to retrieve important and relevant information about data and metadata. This data mining method helps to classify data in different classes.

2. **Clustering**: Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. **Regression**: Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. **Association** **rules**: This data mining technique helps to find the association between two or more items. It discovers a hidden pattern in the data set.

5. **Outer** **detection**: This type of data mining technique refers to observation of data items in the data set which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier Mining.

6.**Sequential patterns**: The data mining technique helps to discover or identify similar patterns.

7. **Prediction** : Prediction has used in the combination of other techniques like trends, clustering, classification etc. It analysis past events and predicts the future event.

**Data mining Examples**

**Example** Consider a marketing head of telecom service provides increase revenues of long distance services. For high ROI on his sales and marketing efforts,customer profiting is important. He has a vast data pool of customer information like age. gender. income.. etc. But its impossible to determine characteristics of people who prefer long distances calls with manual analysis. Using data mining techniques, he may uncover patterns between hig long distance call users and their characteristics.

For example, he might learn that his best customers are married females between the age of 45 and 54 who make more than $80,000 per year. Marketing efforts can be targeted to such demographic.

Example 2: A bank wants to search new ways to increase revenues from its credit card operations. They want to check whether usage would double if fees were halved.

Bank has multiple years of record on average credit card balances, payment amounts, credit limit usage and other key parameters. They create a model to check the impact of the proposed new business policy. The data results show that cutting fees in half for a targeted customer base could increase revenues by $10 million.

**Data Mining Tools**

Following are the two popular Data Mining Tools widely used in the industry:

1. **R Language**: R Language is an open source tool for statistical computing and graphics. R has a wide variety of statistical, classical statistical tests, time-series analysis, classification and graphical techniques. It offers effective data handling and storage facility.

2. **Oracle Data Mining:** Oracle Data Mining, popularly known as ODM, is a module of the Oracle Advanced Analytics Database. This data mining tool allows data analysts to generate detailed insights and makes predictions. It helps predict customer behavior, develops customer profiles and identifies cross-selling opportunities.

**Data Mining — Issues**

Dats mining is not an easy task, as the algorithms used can get very complex and data in not,

always available at one place. It needs to be integrated from vanous heterogencous data sources,

These factors alwo create some issues . major issues are.

1. Mining Methodology and User Interaction lasues

2. Performance Issues

3. Diverse Data Types Issues

**Data Mining System Classification**

A data mining system can be classified according to the following criteria:

• Database Technology

• Statistics

• Machine Learning

• Information Science

• Visualization

• Other Disciplines

**Data Exploration and Reduction**

Data exploration is an approach similar to initial data analysis. whereby a data analyst uses visual exploration to understand what is in a data set and the characteristics of the data, rather than through traditional data management systems.

**Steps of Data Exploration and Preparation**

Remember the quality of your inputs decide the quality of your output. So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here. Data exploration, cleaning and preparation can take up to 70% of your total project time.

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification

2. Univariate Analysis

3. Bi-variate Analysis

4. Missing Values Treatment

5. Outlier Treatment

6. Variable Transformation

7. Variable Creation

Variable Identification

First, identify Predictor (Input) and Target (output) variables. Next. identify the data type and category of the variables.

**Univariate Analysis**

At this stage, we explore variables one by one. Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.

Continuous Variables:

In case of continuous variables, we need to understand the central tendency and spread of the variable.

Categorical Variables:

For categorical variables, we will use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

**Bivariate Analysis**

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical and Categorical, Categorical and Continuous, and Continuous and Continuous. Different methods are used to tackle these combinations during analysis process.

Let's understand the possible combinations in detail.

Continuous and Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatterplot. It is a nifty way to find out the relationship between two variables. The pattern of scatterplot indicates the relationship between variables. The relationship can be linear or non-linear.

Scatterplot shows the relationship between two variable but does not indicate the strength of relationship amongst them\_ To find the strength of the relationship, we use Correlation. Correlation varies between —1 and +1.

Correlation can be derived using following formula:

Correlation = Covariance(X,Y) / SQRT(Var(X)\* Var(Y))

Categorical and Categorical: To find the relationship between two categorical variables, we can use following methods:

Two-way table: We can start analyzing the relationship by creating a two-way table of Count and Count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show Count or Count% of observations available in each combination of row and column categories.

Stacked Column Chart: This method is more of a visual form of two-way table.

Chi-Square Test:

This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-Square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed Chi-Square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable are dependent.

Probability of 1: It shows that both variables are independent Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence.

Categorical and Continuous: While exploring relation between categorical and continuoa3 dra,‘,. box plots for each level of categorical variables. If levels are small in number, chow the statistical significance. To look at the statistical significance, we can perform Z-test or ttest or ANOVA\_

Z-test' /Ttest: Either test assess whether mean of two groups are statistically different from each other or not If the probability of Z is small, then the difference of two averages is more significant The T-test is very similar to Z-test but it is used when number of observations for both cateeories is less than 30.

ANOVA: It assesses whether the average of more than two groups is statistically different For example, suppose we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks. We need to find out whether the effect of these exercises on them is significantly different or not This can be done by comparing the weights of the 5 groups of 4 men each.

Data Reduction

Data reduction is the process of minimizing the amount of data that needs to be stored in a storage environement.It can increase storage efficiency and reduce costs.

Dimensionality and Data Reduction Assume that the data to be reduced consists of tuples or data vectors described by characteristics. Principal components analysis or PCA or Klmethod, searched for k dimensional vectors that can be used to interpret the data.

**Data reduction techniques** can be applied to obtain a compressed representation of the data set that is much smaller in volume, yet maintains the integrity of the original data. Strategies for data reduction include:

Data cube aggregation: Aggregation operations are applied to the data in the construction of a data cube.

Attribute subset selection: Irrelevant, weakly relevant or redundant characteristics or dimensions may be detected and removed.

Dimensionality reduction: Encoding mechanisms are used to reduce the data set size.

Numerosity reduction: Data is replaced or estimated by alternative, smaller data representations such as parametric models (which need to store only the model parameters instead of the actual data) or non-parametric methods such as clustering, sampling and the use of histograms.

Discretization and concept hierarchy generation: Raw data values for attributes are replaced by series or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction. The computational time spent on data reduction should not outweigh or erase the time saved by mining on a reduced data set size.

**What are Data Cubes?**

Data cubes store multidimensional aggregated information. Each cell holds an aggregate data value, corresponding to the data point in multidimensional space. Concept hierarchies may exist for each attribute, allowing the analysis of data at multiple levels of abstraction. For example, a hierarchy for branch could allow branches to be grouped into regions, based on their location.

**Data Classification**

Classification models predict categorical class labels. Following arc the examples of cases where the data analysis task is Classification:

• A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which arc safe.

• A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

How Does Classification Works?

With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps:

1. Building the Classifier or Model

2. Using Classifier for Classification Building the Classifier or Model

• This step is the learning step or the learning phase.

• In this step, the classification algorithms build the classifier.

• The classifier is built from the training set made up of database tuples and their associated class labels.

• Each topic that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

**Classification Issues**

The major issue is preparing the data for Classification and Prediction. Preparing the data involves the following activities:

Data Cleaning: Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

Relevance Analysis: Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes arc related.

Data Transformation and Reduction: The data can be transformed by any of the following methods:

Normalization: The data is transformed using normalization. It involves scaling all Values for given attribute in order to make them fall within a small specified range. It is used when in the learning step, the neural networks or the methods involving measurements are used.

Generalization: The data can also be transformed by generalizing it to the highci concept. For this purpose, we can use the concept hierarchies. Note: Data can also be reduced by some other methods such us wavelet transformation, binning, histogram analysis and clustering.

**Data Association**

Association rules are if-then statements that help to show the probability of relationships between data items within large data sets in various types of databases. Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

**Now Association Rules Work?**

Association rule mining, at a basic level, involves the use of machine learning models to analyze data for patterns, or co-occurrence, in a database. It identifies frequent if-then associations, which are called association rules.

An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent. Association rules are created by searching data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships.

Association rules are calculated from item sets, which are made up of two or more items. If rules

are built from analyzing all the possible item sets, there could be so many rules that the rules hold little meaning. With that, association rules are typically created from rules well-represented in data.

**Association Rule Algorithms**

Popular algorithms that use association rules include AIS, SETM, a priori and variations of the

latter.

With the AIS algorithm, item sces are generated and counted as it scans the data. In transaction

data, the AIS algorithm determines which large item sets contained a transaction, and new candidate

item sets are created by extending the large item sets with other items in the transaction data.

The SETM algorithm also gencrates candidate item sets as it scans a database, but this algorithm

accounts for the item sets at the end of its scan. New candidate item scts arc generated the same way

as with the AIS algorithm, but the transaction ID of the generating transaction is saved with the

candidate item set in a sequential structure. At the end of the pass, the support count of candidate item sets is created by aggregating the sequential structure.

The downside of both the AIS and SETM algorithms is that each one can generate and count

many small candidate item sets, according to published materials from Dr. Sacd Sayad, author of

“Real-time Data Mining”.

**Uses of Association Rules in Data Mining**

In data mining, association rules are useful for analyzing and predicting customer behavior. They

play an important part in customer analytics, market basket analysis, product clustering, and catalog

design and store layout.

**Examples of Association Rules In Data Mining**

A classic example of association rule mining refers to a relationship between diapers and beers.

The example, which seems to be fictional, claims that men who go to a store to buy diapers are also

likely to buy beer. Data that would point to that might look like this:

A supermarket has 200,000 customer transactions. About 4,000 transactions, or about 2% of total

transactions, include the purchase of diapers. About 5,500 transactions (2.75%) include the purchase

of beer. Of those, about 3,500 transactions, 1.75%, include both the purchase of diapers and beer.

Based on the percentages, that number should be much lower. However, the fact that about 87.5% of diaper purchases include the purchase of beer indicates a link between diapers and beer.

**Cause and Effect Modeling in Data Mining**

Cause and effect refers to a relationship between two phenomena in which one phenomenon ;

the reason behind the other. For example, cating too much fast food without any physical activity leads to weight gain. Here, eating without any physical activity is the “cause” and weight gain is th

“effect.” Another popular cxample in the discussion of cause and effect is that of smoking and lune

cancer. A question that has surfaced in cancer research in the past several decades is, what is the offen of smoking on an individual's health?

There are four steps to using the tool:

1. Identify the problem.

2. Work out the major factors involved.

3. Identify possible causes.

4. Analyze your diagram.

**Unit 4**

Prescriptive Analytics is the area of business analytics (BA) dedicated to finding the best course

of action for a given situation. Prescriptive analytics is related to both descriptive and predictive

analytics.

While descriptive analytics aims to provide insight into what has happened and predictive

analytics helps model and forecast what might happen, prescriptive analytics seeks to determine the best solution or outcome among various choices, given the known parameters.

Prescriptive analytics can also suggest decision options for how to take advantage of a future

opportunity or mitigate a future risk, and illustrate the implications of each decision option. In practice,prescriptive aes continually and automatically process new data to improve the accuracy of

redictions and provide better decision options.

**What is Linear Programming (Linear Optimization)?**

What is linear programming? Linear programming is a simple technique where we depict

complex relationships through linear functions and then find the optimum points. The important word in previous sentence is depict. The real relationships might be much more complex, but we can

simplify them to linear relationships.

Applications of linear programming are everywhere around you. You use linear programming at

personal and professional fronts. You are using linear programming when you are driving from home

to work and want to take the shortest route. Or when you have a project delivery, you make strategies

to make your team work efficiently for on-time delivery.

**Example of a Linear Programming Problem**

Let's say a FedEx delivery man has 6 packages to deliver in a day. The warehouse is located at

point A. The 6 delivery destinations are given by U, V, W, X, Y and Z. The numbers on the lines

indicate the distance between the cities. To save on fuel and time, the delivery person wants to take theshortest route.

So, the delivery person will calculate different routes for going to all the 6 destinations and the

come up with the shortest route. This technique of choosing the shortest route is called linea

programming.

In this case, the objective of the delivery person is to deliver the parcel on time at ql] 6

destinations. The process of choosing the best route is called Operations Research. Operation,

research is an approach to decision-making, which involves a set of methods to operate a system,

the above example, my system was the Delivery model.

Lincar Programming is used for obtaining the most optimal solution for 4 problem with Riven

constraints. In lincar programming, we formulate our real-life problem into a mathematical Model, It

involves an objective function, linear inequalities with subject to constraints.

Is the linear representation of the 6 points above representative of real world? Yes and No. It is

oversimplification as the real route would not be a straight line. It would likely have multiple turns, signals and traffic jams. But with a simple assumption, we have reduced the complexity Of the

problem drastically and are creating a solution which should work in most scenarios.

Common Terminologles Used In Linear Programming

Let us define some terminologies used in Linear Programming using the above example.

Decision Variables: The decision variables are the variables which will decide the output.

They represent the ultimate solution. To solve any problem, we first need to identify the

decision variables. For the above example, the total number of units for A and B denoted by

X and Y respectively are the decision variables.

Objective Function: It is defined as the objective of making decisions. In the above

example, the company wishes to increase the total profit represented by Z. So, profit is the

objective function.

Constraints: The constraints are the restrictions or limitations on the decision variables.

They usually limit the value of the decision variables. In the above example, the limit on the

availability of resources Milk and Choco are the constraints.

Non-negativity restriction: For all linear programs, the decision variables should always

take non-negative values which means the values for decision variables should be greater

than or equal to 0.

**Non Linear Programming**

Non Linear Programming is the process of solving an optimisation problem, where some of the constraints or the objective function are non-linear.

In the substitution method, the constraint equation is solved for one variable in terms of another

and then substituted into the Objective function.

An Example - The Transportation Problem with Volume Discounts

Determine an optimal plan for shipping goods from various sources to various destinations,

given supply and demand constraints.

In actuality, the shipping costs may not be fixed. Volume discounts sometimes are available |

for large shipments, which cause a piece-wise linear cost function.

**Cutting Plane Algorithm and Other Methods**

**Cutting Plane Methods**

Consider a pure integer linear Programming problem in which all parameters are integer. This can

be accomplished by multiplying the constraint by a suitable constant. Because of this assumption also, the objective function value and all the “slack” variables of the problem must have integer values.

Cutting Plane Algorithm

Solve the LP relaxation.

stop, if all variables in the solution have integer values: then it is the optimum. Stop, if the

problem is infeasible or unbounded.

1.Select a row with a fractional RHS: this is called the source row. Generate the cut constraint

associated with this row:

2, Augment the simplex tableau with a column for s; and the new constraint row.

3. Solve the LP problem with the dual simplex method.

Stop, if all variables in the solution have integer values, then it is the optimum. Stop, if the

problem is infeasible or unbounded.

**Decision-making under Certainty**

A condition of certainty exists when the decision-maker knows with reasonable certainty what

alternatives are, what conditions are associated with each alternative, and the outcome of each

alternative. Under conditions of certainty, accurate, measurable and reliable information on which to

have decisions is available.

The cause and effect relationships are known and the future is highly predictable under conditions

of certainty. Such conditions exist in case of routine and repetitive decisions concerning the day-to-

day operations of the business.

**Decision-making under Risk**

When a manager lacks perfect information or whenever an information asymmetry exists, risk

arises. Under a state of risk, the decision-maker has incomplete information about available

alternatives but has a good idea of the probability of outcomes for each alternative.

While making decisions under a state of risk, managers must determine the probability associated

with each alternative on the basis of the available information and his experience.

**Decision-making under Uncertainty**

Most significant decisions made in today’s complex environment are formulated under a stat

uncertainty. Conditions of uncertainty exist when the future environment: is unpredictable

everything is ina state of flux. The decision-maker is not aware of all available alternatives, the r

associated with cach, and the consequences of each altemative or their probabilities.

**Modern Approaches to Decision-making under Uncertainty**

There are several modem techniques to improve the quality of decision-making under Conditions

of uncertainty.

**Risk Analysis**

Managers who follow this approach analyze the size and nature of the risk involved in choosing a

particular course of action.

For instance, while launching a new product, a manager has to carefully analyze cach of the

following variables - the cost of launching the product, its pen@urstan: Sees “ capital investment

required, the price that can be set for the product, the potential market size and what per cent Of the

total market it will represent. .

Risk analysis involves quantitative and qualitative risk assessment, risk management and risk

communication and provides managers with a better understanding of the risk and the benefits

associated with a proposed course of action. The decision represents a trade-off between the risks and

the benefits associated with a particular course of action under conditions of uncertainty.

**Decision Trees**

These are considered to be one of the best ways to analyze a decision. A decision tree approach

involves a graphic representation of alternative courses of action, and the possible outcomes and Tasks associated with each action.

By means of a “tree” diagram depicting the decision points, chance events and Probabilities

involved in various courses of action, this technique of decision-making allows the decision-maker to

trace the optimum path or course of action.

**Preference or Utility Theory**

This is another approach to decision-making under conditions of uncertainty. This approach is

based on the notion that individual attitudes towards risk vary. Some individuals are willing to take

only smaller risks (“risk averters”), while others are willing to take greater risks (“gamblers”),

Statistical probabilities associated with the various courses of action are based on the assumption that

decision-makers will follow them.

For instance, if there were a 60° chance of a decision being nght, it might seem reasonable that a

person would take the risk. This may not be necessarily true as the individual might not wish to take

the risk, since the chances of decision being wrong are 40%.

Top-level managers usually take the largest amount of risk. However, the same managers who make

a decision that risks millions of rupees of the company in a given program with a 75'% chance o

mee es are not likely to do the same with their own money,

Moreover, a manager willing to take a 75% risk in one situation may not be willing to do the another. Similarly, a top executive might launch an advertising campaign having a 70% chance of success but might decide against investing in plant and machinery unless it involves a higher probability of success.

‘Though personal attitudes towards risk vary, two things arc certain.

1. Attitudes towards risk vary with situations, i.e, some people are risk averters in some

‘ situations and gamblers in others.

2. Some people have a high aversion to risk, while others have a low aversion.

**Unit 5**

R Environment :

R is a programming language and software environment for statistical analysis, graphics

representation and reporting.

**Evolution of R**

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the

University of Auckland in Auckland, New Zealand. R made its first appearance in 1993.

A large group of individuals has contributed to R by sending code and bug reports.Since mid-1997, there has been a core group (the “R Core Team”) who can modify the R source ~

code archive.

**Features of R**

As stated earlier, R is a programming language and software environment for statistical analysis,

graphics representation and reporting. The following are the important features of R:

R is a well-developed, simple and effective programming language which includes

conditionals, loops, user-defined recursive functions and input-output facilities.

R has an effective data handling and storage facility.

R provides a suite of operators for calculations on arrays, lists, vectors and matrices

R provides a large, coherent and integrated collection of tools for data analysis. a

R provides graphical facilities for data analysis and display either directly at the computer or a

printing at the papers.

environment Setup

If you are still willing to sct up your environment for R, you can follow the steps given below.

“a Installation can download the Windows installer version of R from R-3.2.2 for Windows (32/64 bit) and in’ local directory.

It is a Windows installer (.exe) with a name “R-version-win.exe”, you can just double click

the installer accepting the default settings. If your Windows is 32-bit version, it installs the 32.

gd ea But if your windows is 64-bit, then it installs both the 32-bit and 64-bit versions.

**R command prompt**

Once you have R environment setup, then it is easy to start your R command prompt by just

ing the following command at your command prompt .

This will launch R interpreter and you will get a prompt > where you can start typing your

program as follows:

> myString <- "Hello, World!"

> print ( myString)

(1) "Hello, World!”

Here, the first statement defines a string variable myString, where we assign a string "Hello,

world!" and then next statement print() is being used to print the value stored in variable myString.

**R Script File**

R interpreter called Rscript. So, let’s start with writing following code in a text file called testR

as under:

# My first program in R Programming

myString <- "Hello, World!" -

print (myString)

Save the above code in a file test.R and execute it at Linux command prompt as given below.

Even if you are using Windows or other system, the syntax will remain same.

**Data Types**

There are many types of R-objects, The frequently used ones are:

1. Vectors

2. Lists

3. Matrices

4. Arrays

5. Factors

6. Data Frames

**Vectors**

When you want to create vector with more than one clement, you should use c() function which

means to combine the elements into a vector.

# Create a vector.

apple <- c(‘red',’green’,"yellow")

print(apple)

# Get the class of the vector.

print(class(apple))

When we execute the above code, it produces the following result:

[1] "red" “green” "yellow"

[1] “character”

**Lists**

A list is an R-object which can contain many different types of elements inside it like vectors,

functions and even another list inside it.

# Create a list.

list1 <- list(c(2,5,3),21.3,sin)

# Print the list

print(list1)

When we execute the above code, it produces the following result:

253

21.3

**Matrix**

A Matrix Is a two-dimensional Rectangular data set. It can be created using vector input to matrix function.

M= matrix( c(‘a’,’b’,’c’,’d’,’e’,’f’), nrow = 2, ncol = 3, byrow = TRUE)

**Arrays**

While matrices are confined to two dimensions, arrays can be of any number of dimensions The

array function takes a dim attribute which creates the required number of dimension. In the below

example, we create an array with two clements which are 3 « 3 matrices cach.

# Create an array.

a <- array(c(‘green’,'yellow’),dim = c{3,3,2))

print(a)

**Factors**

Factors are the R-objects which are created using a vector: It stores the vector along with the

distinct values of the elements in the vector as labels. The labels are always character irrespective of

whether it is numeric or character or Boolean, etc. in the input vector. They are useful in statistical

modeling.

Factors are created using the factor() function. The nlevels functions gives the count of levels.

# Create a vector.

apple\_colors <- c(‘green’,'green','yellow’,‘red’ ‘red’ red’, green’)

# Create a factor object.

factor\_apple <- factor(apple\_colors)

# Print the factor.

print(factor\_apple)

print(nlevels(factor\_apple))

**Data Frames**

Data frames are tabular data objects. Unlike a matrix in data frame, each column can contain

different modes of data. The first column can be numeric while the second column can be character

and third column can be logical. It is a list of vectors of equal length.Data Frames are created using the data.frame() function.

# Create the data frame.

BMI <- data.frame(

gender = c("Male", "Male","Female\*),

height = c(152, 171.5, 165).

weight =c (81,93, 78),

Age = c(42,38,26)

)

print(BMI)

**Types of Operators**

We have the following types of operators in R Programming:

1. Anthmetic Operators

2. Relational Operators

3. Logical Operators

4. Assignment Operators

5. Miscellaneous Operators

**R Variables**

A variable provides us with named storage that our programs can manipulate. A variable in R can

store an atomic vector, group of atomic vectors or a combination of many objects. A valid variable

name consists of letters, numbers and the dot or underline characters. The variable name starts with a

letter or the dot not followed by a number.

**R Packages**

R Packages are a collection of R functions, complied code and sample data. They are stored

under a directory called “library” “in the R environment. By default, R installs a set of packages

during installation. More packages are added later, when they are needed for some specific purpose.

When we start the R console, only the default packages are available by default. Other packages which are already installed have to be loaded explicitly to be used by the R program that is going to use them.

**Install a New Package**

There are two ways to add new R packages — one is installing directly from the CRAN directory

and another is downloading the package to your local system and installing it manually.

1. Install directly from CRAN

The following command gets the packages directly from CRAN webpage and installs the

package in the R environment. You may be prompted to choose a nearest mirror. Choose the

one appropriate to your location.

install.packages("Package Name")

# Install the package named "XML".

install.packages("XML")

2. Install package manually

Go to the link R Packages to download the package needed. Save the package as a .zip file

in a suitable location in the local system.

Now you can run the following command to install this package in the R environment.

**Reading and Writing Datain R**

**Functions for Reading Data into R:**

There are a few very useful functions for reading data into R.

read.table() andread.csv() are two popular functions used for reading tabular data into R.

readLines() is used for reading lines from a text file.

source{) is a very useful function for reading in R code files from a another R program.

dget() function is also used for reading in R code files.

load() function is used for reading in saved workspaces

unserialize() function is used for reading single R objects in binary format.

**Functions for Writing Data to Files:**

There are similar functions for writing data to files

write.tableQ) is used for writing tabular data to text files (ie. CSV).

writeLines() function is useful for writing character data line-by-line to a file or connection.

dump( is a function for dumping a textual representation of multiple R objects.

dput() function is used for outputting a textual representation of an R object.

save() is useful for saving an arbitrary number of R objects in binary formatto a file.

serialize() is used for converting an R object into a binary format for outputting to a

connection (or file).

**R Functions**

A function is a set of statements organized together to perform a specific task. R has a large

qumber of in-built functions and the user can create their own functions.

In R, a function is an object. So, the R interpreter is able to pass control to the function along with

arguments that may be necessary for the function to accomplish the actions.

The function in tum performs its task and returns control to the interpreter as well as any result

which may be stored in other objects.

**Function Definition**

An R function is created by using the keyword function. The basic syntax of an R function

definition is as follows:

function\_name <- function(arg\_1, arg\_2, ...) {

Function body

}

**Function Components**

The different parts of a function are:

Function Name: This is the actual name of the function. It is stored in R environment as an

object with this name.

Arguments: An argument is a placeholder. When a function is invoked, you can pass a

value to the argument. Arguments are optional, i.e., a function may contain no arguments.

Also, arguments can have default values.

Function Body: The function body contains a collection of statements that defines what the

function does.

Return Value: The return value of a function is the last expression in the function body to

be evaluated.

In-built and user-defined functions: R has many in-built functions which can be directly

called in the program without defining them first We can also create and use our own

functions referred as user-defined functions.

**User-defined Function**

We can create user-defined functions in R. They are specific to what a user wants,

created, they can be used like the built-in functions. Below is an example of how a function j

and used.

# Create a function to Print squares of numbers in sequence.

new. function <- function(a) {

for(i in 1:a) {

b<-K2

Print(b)

}

)

Calling a Function

# Create a function to print squares of numbers in sequence.

new. Function <- function (a) {

for (I in 1:a) (

b<- #2

Print (b)

)

)

# Call the function new. Function supplying 6 as an argument.

new. Function (6)

**Control Statements**

**Decision-making Statements in R**

Decision-making structures require the programmer to specify one or more conditions to be

evaluated or tested by the program, along with a statement or statements to be executed if the

condition is determined to be true, and optionally, other statements to be executed if the condition is

determined to be false.

Following is the general form of a typical decision-making structure found in most of the

programming languages.

if statement: An if statement consists of a Boolean expression followed by one or more

statements.

if..else statement: An if statement can be followed by an optional else statement, which

executes when the Boolean expression is false.

switch statement: A switch statement allows a variable to be tested for equality against a

list of values.

**Loops inR**

There may be a situation when you need to execute a block of code several number of times. In

general, statements are executed sequentially. The first statement in a function is executed first,

followed by the second, and so on.

Programming languages provide various control structures that allow for more complicated

execution paths.

A loop statement allows us to execute a statement or group of statements multiple times. The

following is the general form of a loop statement in most of the programming languages:

R programming language provides the following kinds of loop to handle looping requirements.

repeat loop: Executes a sequence of statements multiple times and abbreviates the code that

manages the loop variable.

while loop: Repeats a statement or group of statements while a given condition is true. It

tests the condition before executing the loop body.

for loop: Like a while statement, except that it tests the condition at the end of the loop

body.

**Frames and Subsets**

**Data Frames**

A data frame is a table or a two-dimensional array-like structure in which each column contains

values of one variable and each row contains one set of values from each column.

Following are the characteristics of a data frame.

The column names should be non-empty.

The row names should be unique.

The data stored in a data frame can be of numeric, factor or character type.

Each column should contain same number of data items.

**Subset Data in R**

In R, the command “subset” is used to filter the data in a data frame based on the criteria you set.

The below will show you how subset works and provides some subset examples.

The first example selects the records in the data frame Student Data where Grade is 3 and copies

those records back into the same data frame. Note that the Grade field is numeric.

# Subset data inR

StudentData<-subset (Student Data, Grade= =3)

The problem with the above is that all the records where Grade is not 3 have been lost. In this

example, R selects the records from the data frame Student Data where Grade is 3 and copies those

records to a new data frame Grade3StudentData, preserving all of the records for later use. .

**Set Your Working Directory**

Now, we can set the working directory to this folder by selecting from the top menu Session >

Set Working Directory > To Source File Location. (Doing so means we can load the files in this

directory without having to refer to the full path for their location, and anything we save will be

written to this folder.)

Notice how this code appears in the console:

setwd ("“~/Desktop/week7")

**Save Your Data**

The panel at top right has two tabs, the first showing the Environment, or all of the “objects”

Loaded into memory for this R session. We can save this as well, so we do not have to load and process data again if we retum to retum to a project later.

(The second tab shows the History of the operations you have performed in RStudio.)

Click on the save/disk icon in the Environment panel to save and call the file week7.RData.

You should sce the following code appear in the Console:

[save Image ("~/Desktop/week7/week7 RData") |

Copy this code into your script, placing it at the end, with a comment, explaining what it docs:

# save session data .

Save. Image ("~/Desktop/week7/week7.RData”)

**Comment Your Code**

Anything that appears on a line after # will be treated as a comment, and will be ignored when the

code is run. Get into the habit of commenting your code: “Don’t trust yourself to remember what it

does!”

**Load and View Data**

**Load Data**

You can load data into the current R session by selecting Import Datasct > From Text File... in

the Environment tab.

However, we will use the read\_esv function from the readr package. Copy the following code

into your srnipt and Run:

# load data of Pfizer payments to doctors and warming letters sent by food and dm

administranon 8

Pfizer <- read\_csv (“pfizer.csv\*)

fda <- read\_csv (“Yda.csv”)

**Filter and Sort Data**

Now, we will filter and sort the data in specific ways. For each of the following examples, copy

the code that follows into your script and view the results. Notice how we create a new objects to hold the processed data.

Find doctors in California paid $10,000 or more by Pfizer to run “Expert-Led Forums.”

# doctors in California who were paid $10,000 or more by Pfizer to run “Expert-Led

Forums.”

ca\_expert\_10000 <- pfizer %>%

filter (state = = "CA" & total >= 10000 & category = = "Expert-Led Forums")

Notice the use of = = to find values that match the specified text, >= for greater than or equal to,

and the Boolean operator &.

**Join data in R**

There are also a number of join functions in dplyr to combine data from two data frames, Here

are the most useful:

Inner\_Join (: Retums values from both tables only where there is a match

Ieft\_join (: Retums all the valucs from the first-mentioned table, plus those from the

second table that match.

semi\_join (): Filters the first-mentioned table to include only values that have matches in

the second table,

anti\_join (): Filters the first-mentioned table to include only values that have no matches jn

the second table.