ABES Engineering College, Ghaziabad
B. Tech Odd Semester Sessional Test-1

Printed Pages:
Session: 2023-24

Semester: 5th
Course Code: KDS501 **Roll No.:**
Course Name: Data Analytics & Visualization **Time: 1.15 Hrs.**
Maximum Marks: 30
Instructions:

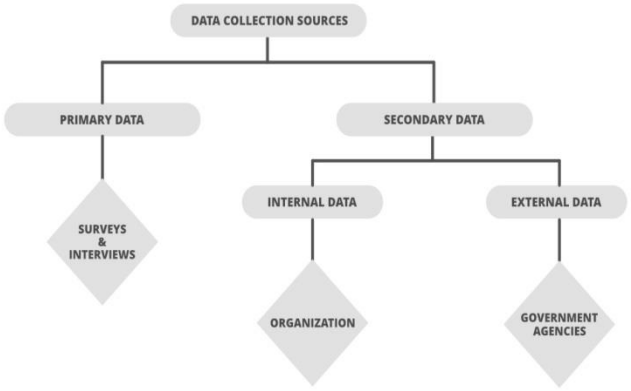
1. **Attempt All sections.**
2. **If require any missing data, then choose suitably.**

Q. No	Question	Marks	CO	KL	PI
	Section-A	**Total Marks : 20**			
1	**Attempt ANY ONE part from the following**	***Same K Levels Questions***			
a)	**Elaborate different types of Data Analytics used in Business Decision Making and present their importance accordingly**. **Ans: 1. Descriptive:** If you have a question about something that has already happened, then descriptive analytics can help you answer it. Descriptive analytics is often used as a means to explain something to stakeholders. For example, it can track return on investment (ROI) and other metrics of past performance. As useful as descriptive analytics is, it's best to combine descriptive analytics with another method like diagnostic to go deeper into why something has happened. Descriptive analytics will point out what happened, but you will need to explore the reasoning behind the event still. **2. Diagnostic:** Like a diagnosis, diagnostic analytics provides insight as to why something happened. They work hand-in-hand with descriptive analytics to further explain critical findings. ▪ If you take a look at key performance indicators (KPIs) and want to understand why something improved or got worse, then diagnostic analytics help to: ▪ Identify anomalies in data ▪ Collect the data that helps to understand the changes ▪ Uses statistical techniques to help explain such anomalies **3. Predictive:** Answer questions about what could happen in the future. This analytic method leverages past data to evaluate trends and estimate the likelihood of something recurring. Statistical analysis, regression and machine learning is used to make predictive analytics function. **4. Prescriptive:** If you find yourself in a critical position to make a decision about the future but feel unsure about what choice to make, prescriptive analytics can be a lifesaver. Prescriptive analytics works by finding patterns from large datasets and then estimates the likelihood of different outcomes. **5. Cognitive:** Cognitive Analytics applies human-like intelligence to certain tasks, and brings together a number of intelligent technologies, including semantics, artificial intelligence algorithms, deep learning and machine learning.	**5**	**CO 1**	K2	2.1.3

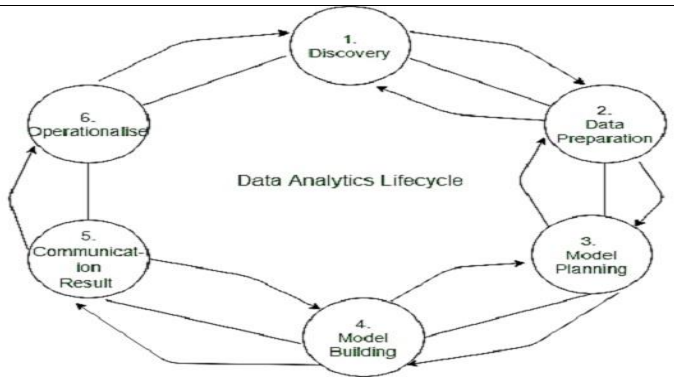
b)	**Data plays a vital role for effective data analysis. For this discuss the nature of data and various characteristics of it using suitable example.** **Ans: Nature of Data** 1. Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. • **Nominal Data:** ➢ Nominal Data is used to label variables without any order or quantitative value. The colour of hair can be considered nominal data, as one colour can't be compared with another colour. ➢ The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed to distinct categories. • **Ordinal Data** ➢ Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them. 2. Quantitative data is the value of data in the form of counts or numbers where each data set has a unique numerical value. This data is any quantifiable information that researchers can use for mathematical calculations and statistical analysis to make real-life decisions based on these mathematical derivations. ▪ **Discrete Data** ➢ The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table. ▪ **Continuous Data** ➢ Continuous data are in the form of fractional numbers. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range. ➢ The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc. **Characteristics of Data** 1. **<u>Accuracy</u>:** As the name implies, this data quality characteristic means that information is correct. To determine whether data is accurate or not, ask you if the information reflects a real-world situation. Accuracy is a crucial data quality characteristic because inaccurate information can cause significant problems with severe consequences. 2. **<u>Completeness</u>**: "Completeness" refers to how comprehensive the information is. When looking at data completeness, think about whether all of the data you need is available; you might need a customer's first and last name, but the middle initial may be optional. 3. **<u>Reliability</u>**: Reliability means that a piece of information doesn't contradict another piece of information in a different source or system.	**5**	**CO 1**	K2	2.2.3

2	**Attempt ANY ONE part from the following**	***Same K Levels Questions***			
a)	**Introduce Big Data Analytics and various challenges come along with big data. Big data can be defined by 5 V's described them in detail.** **Ans: Big Data:** Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size. **Types of Big Data** ▪ Following are the types of Big Data: ▪ Structured ▪ Unstructured ▪ Semi-structured **5 V's of Big Data** Big Data was defined by the "*3Vs*" but now there is "*5Vs*" of Big Data which are also termed as the characteristics of Big Data as follows: 1. **Volume:** The name 'Big Data' itself is related to a size which is enormous. Volume is a huge amount of data. To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. 2. **Velocity:** Velocity refers to the high speed of accumulation of data. In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc. There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands. Sampling data can help in dealing with the issue like 'velocity'. 3. **Variety:** It refers to nature of data that is structured, semi-structured and unstructured data. It also refers to heterogeneous sources. Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured. 4. **Veracity:** It refers to inconsistencies and uncertainty in data that is data which is available can sometimes get messy and quality and accuracy are difficult to control. Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources. 5. **Value:** After having the 4 V's into account there comes one more V which stands for Value. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful. Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value is the most important V of all the 5V's.	**5**	**CO 1**	K2	2.2.4

b)	**Nowadays businesses are using Big Data for competitive advantages. Justify the importance of big data in different applications**. In today's world, there are a lot of data. Big companies utilize those data for their business growth. By analyzing this data, the useful decision can be made in various cases as discussed below: 1. Tracking Customer Spending Habit, Shopping Behavior: In big retails store (like Amazon, Walmart, Big Bazar etc.) management team has to keep data of customer's spending habit (in which product customer spent, in which brand they wish to spent, how frequently they spent), shopping behavior, customer's most liked product (so that they can keep those products in the store). Which product is being searched/sold most, based on that data, production/collection rate of that product get fixed. Banking sector uses their customer's spending behavior-related data so that they can provide the offer to a particular customer to buy his particular liked product by using bank's credit or debit card with discount or cashback. By this way, they can send the right offer to the right person at the right time. **3.** Recommendation: By tracking customer spending habit, shopping behavior, Big retails store provide a recommendation to the customer. E-commerce site like Amazon, Walmart, Flipkart does product recommendation. They track what product a customer is searching, based on that data they recommend that type of product to that customer. 3. Smart Traffic System: Data about the condition of the traffic of different road, collected through camera kept beside the road, at entry and exit point of the city, GPS device placed in the vehicle (Ola, Uber cab, etc.). All such data are analyzed and jam-free or less jam way, less time taking ways are recommended. Such a way smart traffic system can be built in the city by Big data analysis. One more profit is fuel consumption can be reduced. 4. Secure Air Traffic System: At various places of flight (like propeller etc) sensors present. These sensors capture data like the speed of flight, moisture, temperature, other environmental condition. Based on such data analysis, an environmental parameter within flight are set up and varied. By analyzing flight's machine-generated data, it can be estimated how long the machine can operate flawlessly when it to be replaced/repaired. 5. Auto Driving Car: Big data analysis helps drive a car without human interpretation. In the various spot of car camera, a sensor placed, that gather data like the size of the surrounding car, obstacle, distance from those, etc. These data are being analyzed, then various calculation like how many angles to rotate, what should be speed, when to stop, etc carried out. These calculations help to take action automatically. 6. Virtual Personal Assistant Tool: Big data analysis helps virtual personal assistant tool (like Siri in Apple Device, Cortana in Windows, Google Assistant in Android) to provide the answer of the various question asked by users. This tool tracks the location of the user, their local time, season, other data related to question asked, etc. Analyzing all such data, it provides an answer. As an example, suppose one user asks "Do I need to take Umbrella?", the tool collects data like location of the user, season and weather condition at that location, then analyze these data to conclude if there is a chance of raining, then provide the answer.	**5**	**CO 1**	K2	2.1.3

3	**Attempt ANY ONE part from the following**	***Same K Levels Questions***			
a)	Provide the detailed description of various methods for data collection. Explain "direct Personal Investigation Method" of Collecting Primary Data. Also, Discuss its Merits and Demerits.  ▪ **Primary Data** ▪ Primary data is the data that is collected for the first time through personal experiences or evidence, particularly for research. It is also described as raw data or first-hand information. The mode of assembling the information is costly, as the analysis is done by an agency or an external organisation, and needs human resources and investment. The investigator supervises and controls the data collection process directly. The data is mostly collected through observations, physical testing, mailed questionnaires, surveys, personal interviews, telephonic interviews, case studies, and focus groups, etc. ▪ **Secondary Data** ▪ Secondary data is a second-hand data that is already collected and recorded by some researchers for their purpose, and not for the current research problem. It is accessible in the form of data collected from different sources such as government publications, censuses, internal records of the organisation, books, journal articles, websites and reports, etc. This method of gathering data is affordable, readily available, and saves cost and time. However, the one disadvantage is that the information assembled is for some other purpose and may not meet the present research purpose or may not be accurate. **Method of Collecting Primary Data** **.** Interview method: ▪ The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc. 2. Survey method:	**10**	**CO 1**	K2	2.2.3

	▪ The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls. Observation method: ▪ The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing. ▪ 4. Experimental method: ▪ The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD. ▪ CRD- Completely Randomized design is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments. ▪ RBD- Randomized Block Design is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector. ▪ LSD – Latin Square Design is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. ▪ FD- Factorial design is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.				
b)	Draw a neat diagram and explain different phases of Data Analytics Life Cycle with real world examples. ▪ The **Data Analytics Lifecycle** is a cyclic process which explains, in six stages, how information in made, collected, processed, implemented, and analysed for different objectives.	**10**	**CO 1**	K2	2.2.4



- **1. Data Discovery**
- This is the initial phase to set your project's objectives and find ways to achieve a complete data analytics lifecycle. Start with defining your business domain and ensure you have enough resources (time, technology, data, and people) to achieve your goals.
- The biggest challenge in this phase is to accumulate enough information. You need to draft an analytic plan, which requires some serious leg work.
- **Accumulate resources**
- First, you have to analyze the models you have intended to develop. Then determine how much domain knowledge you need to acquire for fulfilling those models.
- The next important thing to do is assess whether you have enough skills and resources to bring your projects to fruition.
- **Frame the issue**
- Problems are most likely to occur while meeting your client's expectations. Therefore, you need to identify the issues related to the project and explain them to your clients. This process is called "framing." You have to prepare a problem statement explaining the current situation and challenges that can occur in the future. You also need to define the project's objective, including the success and failure criteria for the project.
- **Formulate initial hypothesis**
- Once you gather all the clients' requirements, you have to develop initial hypotheses after exploring the initial data.
- **2. Data Preparation and Processing**
- The Data preparation and processing phase involves collecting, processing, and conditioning data before moving to the model building process.
- **Identify data sources**
- You have to identify various data sources and analyze how much and what kind of data you can accumulate within a given timeframe. Evaluate the data structures, explore their attributes and acquire all the tools needed.
- **Collection of data**
- You can collect data using three methods:
- **Data acquisition:** You can collect data through external sources.
- **Data Entry:** You can prepare data points through digital systems or manual entry as well.

- **Signal reception:** You can accumulate data from digital devices such as IoT devices and control systems.
- **3. Model Planning**
- This is a phase where you have to analyze the quality of data and find a suitable model for your project.
- **Loading Data in Analytics Sandbox**
- An analytics sandbox is a part of data lake architecture that allows you to store and process large amounts of data. It can efficiently process a large range of data such as big data, transactional data, social media data, web data, and many more. It is an environment that allows your analysts to schedule and process data assets using the data tools of their choice. The best part of the analytics sandbox is its agility. It empowers analysts to process data in real-time and get essential information within a short duration.
- **Data are loaded in the sandbox in three ways:**
- **ETL** (Extract, Transform, Load)− Team specialists make the data comply with the business rules before loading it in the sandbox.
- **ELT** − The data is loaded in the sandbox and then transform as per business rules.
- **ETLT** − It comprises two levels of data transformation, including ETL and ELT both.
- The data you have collected may contain unnecessary features or null values. It may come in a form too complex to anticipate. This is where data exploration' can help you uncover the hidden trends in data.
- **Steps involved in data exploration:**
- Data identification
- Univariate Analysis
- Multivariate Analysis
- Filling Null values
- Feature engineering
- For model planning, data analysts often use regression techniques, decision trees, neural networks, etc. Tools mostly used for model planning and execution include Rand PL/R, WEKA, Octave, Statista, and MATLAB.
- **4. Model Building**
- Model building is the process where you have to deploy the planned model in a real-time environment. It allows analysts to solidify their decision-making process by gain in-depth analytical information. This is a repetitive process, as you have to add new features as required by your customers constantly.
- Your aim here is to forecast business decisions and customize market strategies and develop tailor-made customer interests. This can be done by integrating the model into your existing production domain.
- In some cases, a specific model perfectly aligns with the business objectives/ data, and sometimes it requires more than one try. As you start exploring the data, you need to run particular algorithms and compare the outputs with your objectives. In some cases, you may

	even have to run different variances of models simultaneously until you receive the desired results. ▪ **5. Result Communication and Publication** ▪ This is the phase where you have to communicate the data analysis with your clients. It requires several intricate processes where you how to present information to clients in a lucid manner. Your clients don't have enough time to determine which data is essential. Therefore, you must do an impeccable job to grab the attention of your clients. ▪ **Check the data accuracy** ▪ Is the data provide information as expected? If not, then you have to run some other processes to resolve this issue. You need to ensure the data you process provides consistent information. This will help you build a convincing argument while summarizing your findings. ▪ **Highlight important findings** ▪ Well, each data holds a significant role in building an efficient project. However, some data inherits more potent information that can truly serve your audience's benefits. While summarizing your findings, try to categorize data into different key points. ▪ **Determine the most appropriate communication format** ▪ How you communicate your findings tells a lot about you as a professional. We recommend you to go for visuals presentation and animations as it helps you to convey information much faster. However, sometimes you also need to go old-school as well. For instance, your clients may have to carry the findings in physical format. They may also have to pick up certain information and share them with others. **6. Operationalize** ▪ As soon you prepare a detailed report including your key findings, documents, and briefings, your data analytics life cycle almost comes close to the end. The next step remains the measure the effectiveness of your analysis before submitting the final reports to your stakeholders. ▪ In this process, you have to move the sandbox data and run it in a live environment. Then you have to closely monitor the results, ensuring they match with your expected goals. If the findings fit perfectly with your objective, then you can finalize the report. Otherwise, you have to take a step back in your data analytics lifecycle and make some changes.				
Section-B		**Total Marks : 10**			
4	**Attempt ANY ONE part from the following**	***Same K Levels Questions***			
a)	**The Support Vector Machine (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Discuss how this can be overcome and develop a scalable SVM algorithm for efficient SVM classification in large dataset.**	**5**	**CO2**	K 3	1.3.1

Ans: Support Vector Machine (SVM) is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

- To create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyper-plane (MMH) and it can be done in the following two steps −
- First, SVM will generate hyper-planes iteratively that segregates the classes in best way.
- Then, it will choose the hyper-plane that separates the classes correctly.

In machine learning, a kernel refers to a method that allows us to apply linear classifiers to non-linear problems by mapping non-linear data into a higher-dimensional space without the need to visit or understand that higher-dimensional space.

Types of Kernel Function

- **Polynomial Kernel**
- **RBF Kernel**
- **Sigmoid Kernel**

Polynomial Kernel: The polynomial kernel is a kernel function commonly used with support vector machines and other kernelized models, that represents the similarity of vectors in a feature space over polynomials of the original variables. It is popular in image processing.

$$K(x_i, x_j) = (x_i^T . x_j + 1)^d$$

Where d is the degree of polynomial.

T is the transpose.

Radial-Basis Kernel Function: It is general purpose kernel: used when there is no prior knowledge about the data.

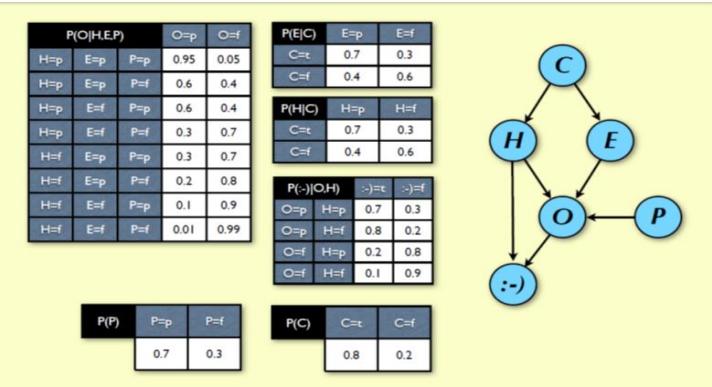
$$K(x, y) = \exp\left(\frac{-|x-y|^2}{2\sigma^2}\right)$$

Hyperbolic Tangent Kernel:

Mainly used in neural networks.

$$K(x_i, x_j) = \tanh\left(kx_i . x_j + c\right)$$

Where k>0 and c<0

b)	There is a test for Allergy to Cats, but this test is not always right. (Assumption population: 1000) • For people that **really do** have the allergy, the test says "Yes" **80%** of the time • For people that **do not** have the allergy, the test says "Yes" **10%** of the time ("false positive") If 1% of the population has the allergy, and **Hunter's test says "Yes"**, what are the chances that Hunter really has the allergy? **<u>Ans:</u>** We want to know the chance of having the allergy when test says "Yes", written **P(Allergy|Yes)** Let's get our formula: P(Allergy|Yes) = *P(Allergy) P(Yes|Allergy)***P(Yes)** • P(Allergy) is Probability of Allergy = 1% • P(Yes|Allergy) is Probability of test saying "Yes" for people with allergy = 80% • P(Yes) is Probability of test saying "Yes" (to anyone) = ??% We **don't know** what the **general** chance of the test saying "Yes" is but we can calculate it by adding up those **with**, and those **without** the allergy: • 1% have the allergy, and the test says "Yes" to 80% of them • 99% do **not** have the allergy and the test says "Yes" to 10% of them Let's add that up: P(Yes) = 1% × 80% + 99% × 10% = 10.7% Which means that about 10.7% of the population will get a "Yes" result. So now we can complete our formula: P(Allergy|Yes) = (*1% × 80%)*/**10.7%** = 7.48% P(Allergy|Yes) = about **7%** **Or** So: **P(A|B)** means "The probability that Hunter actually has the allergy given that the test says Yes" **P(B|A)** means "The probability that the test says Yes given that Hunter actually has the allergy" To be clearer, let's change A to **has** (actually has allergy) and B to **Yes** (test says yes): $P(has|Yes) = \frac{P(has)P(Yes|has)}{P(has)P(Yes|has) + P(not\ has)P(Yes|not\ has)}$ And put in the numbers: $P(has|yes) = \frac{0.01 \times 0.8}{0.01 \times 0.8 + 0.99 \times 0.1} = 0.0748...$ Which is about **7%**	**5**	**CO2**	K 3	1.1.1
5	**Attempt ANY ONE part from the following**			***Same K Levels Questions***	
a)	Explain the steps used in Bayesian Data Analysis. Also calculate the probability of student's happiness depending upon grades using Bayesian inference network.  	**5**	**CO2**	K 3	1.1.1

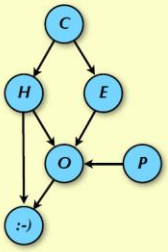
Happiness and Class Attendance

$$\mathbf{P}(\text{:-}), C) = \sum_H \sum_O \sum_P \sum_E \mathbf{P}(C, H, E, O, P, \text{:-}))$$

$$= \sum_{H,O,P,E} Pr(C)\mathbf{P}(P)\mathbf{P}(H \mid C)\mathbf{P}(E \mid C)$$
$$\mathbf{P}(O \mid H, E, P)\mathbf{P}(\text{:-}) \mid H, O)$$

$$= \sum_{H,O,E} Pr(C)\mathbf{P}(H \mid C)\mathbf{P}(E \mid C)\mathbf{P}(\text{:-}) \mid H, O)$$

$$\left(\sum_P \mathbf{P}(P)\mathbf{P}(O \mid H, E, P)\right)$$



This becomes a "function" table $g_1(O,H,E)$.

The Function g_1

$$g_1(O, H, E) = \sum_P \mathbf{P}(P)\mathbf{P}(O \mid H, E, P)$$

P(O|H,E,P)			O=p	O=f
H=p	E=p	P=p	0.95	0.05
H=p	E=p	P=f	0.6	0.4
H=p	E=f	P=p	0.6	0.4
H=p	E=f	P=f	0.3	0.7
H=f	E=p	P=p	0.3	0.7
H=f	E=p	P=f	0.2	0.8
H=f	E=f	P=p	0.1	0.9
H=f	E=f	P=f	0.01	0.99

g_1(O,H,E)			
O=p	H=p	E=p	0.7(0.95)+0.3(0.6)=0.845
O=p	H=p	E=f	0.7(0.6)+0.3(0.3)=0.510
O=p	H=f	E=p	0.7(0.3)+0.3(0.2)=0.270
O=p	H=f	E=f	0.7(0.1)+0.3(0.01)=0.073
O=f	H=p	E=p	0.7(0.05)+0.3(0.4)=0.155
O=f	H=p	E=f	0.7(0.4)+0.3(0.7)=0.490
O=f	H=f	E=p	0.7(0.7)+0.3(0.8)=0.730
O=f	H=f	E=f	0.7(0.9)+0.3(0.99)=0.927

P(P)	P=p	P=f
	0.7	0.3

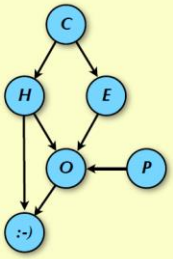
Happiness and Class Attendance

$$\mathbf{P}(\text{:-}), C) = \sum_{H,O,E} Pr(C)\mathbf{P}(H \mid C)\mathbf{P}(E \mid C)\mathbf{P}(\text{:-}) \mid H, O)g_1(O, H, E)$$

$$= \sum_{H,O} Pr(C)\mathbf{P}(H \mid C)\mathbf{P}(\text{:-}) \mid H, O)\left(\sum_E \mathbf{P}(E \mid C)g_1(O, H, E)\right)$$

$$g_2(O, H, C) = \sum_E \mathbf{P}(E \mid C)g_1(O, H, E)$$

Another table to compute



The Function g_2

$$g_2(O, H, C) = \sum_E \mathbf{P}(E \mid C)g_1(O, H, E)$$

g_1(O,H,E)			
O=p	H=p	E=p	0.845
O=p	H=p	E=f	0.510
O=p	H=f	E=p	0.270
O=p	H=f	E=f	0.073
O=f	H=p	E=p	0.155
O=f	H=p	E=f	0.490
O=f	H=f	E=p	0.730
O=f	H=f	E=f	0.927

g_2(O,H,C)			
O=p	H=p	C=t	0.7(0.845)+0.3(0.510)=0.7445
O=p	H=p	C=f	0.4(0.845)+0.6(0.510)=0.644
O=p	H=f	C=t	0.7(0.270)+0.3(0.073)=0.2109
O=p	H=f	C=f	0.4(0.270)+0.6(0.073)=0.1518
O=f	H=p	C=t	0.7(0.155)+0.3(0.490)=0.2555
O=f	H=p	C=f	0.4(0.155)+0.6(0.490)=0.356
O=f	H=f	C=t	0.7(0.730)+0.3(0.927)=0.7891
O=f	H=f	C=f	0.4(0.730)+0.6(0.927)=0.8482

P(E|C)	E=p	E=f
C=t	0.7	0.3
C=f	0.4	0.6

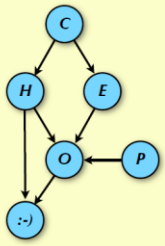
Happiness and Class Attendance

$$\mathbf{P}(\text{:-}), C) = \sum_{H,O} Pr(C)\mathbf{P}(H \mid C)\mathbf{P}(\text{:-}) \mid H, O)g_2(O, H, C)$$

$$= \sum_{H} Pr(C)\mathbf{P}(H \mid C)\left(\sum_{O} \mathbf{P}(\text{:-}) \mid H, O)g_2(O, H, C)\right)$$

$$g_3(\text{:-}), H, C) = \sum_{O} \mathbf{P}(\text{:-}) \mid H, O)g_2(O, H, C)$$

Another table to compute



The Function g_3

$$g_3(\text{:-}), H, C) = \sum_{O} \mathbf{P}(\text{:-}) \mid H, O)g_2(O, H, C)$$

P(:-)|O,H)		:-)=t	:-)=f
O=p	H=p	0.7	0.3
O=p	H=f	0.8	0.2
O=f	H=p	0.2	0.8
O=f	H=f	0.1	0.9

g_2(O,H,C)			
O=p	H=p	C=t	0.7445
O=p	H=p	C=f	0.644
O=p	H=f	C=t	0.2109
O=p	H=f	C=f	0.1518
O=f	H=p	C=t	0.2555
O=f	H=p	C=f	0.356
O=f	H=f	C=t	0.7891
O=f	H=f	C=f	0.8482

g_3(:-),H,C)			
:-)=t	H=p	C=t	0.7(0.7445)+0.2(0.2555)=0.5723
:-)=t	H=p	C=f	0.7(0.644)+0.2(0.356)=0.522
:-)=t	H=f	C=t	0.8(0.2109)+0.1(0.7891)=0.2468
:-)=t	H=f	C=f	0.8(0.1518)+0.1(0.8482)=0.2063
:-)=f	H=p	C=t	0.3(0.7445)+0.8(0.2555)=0.4278
:-)=f	H=p	C=f	0.3(0.644)+0.8(0.356)=0.478
:-)=f	H=f	C=t	0.2(0.2109)+0.9(0.7891)=0.752
:-)=f	H=f	C=f	0.2(0.1518)+0.9(0.8482)=0.793

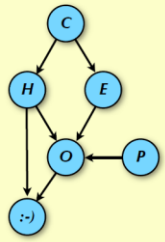
Happiness and Class Attendance

$$\mathbf{P}(\text{:-}), C) = \sum_{H} Pr(C)\mathbf{P}(H \mid C)g_3(\text{:-}), H, C)$$

$$= Pr(C)\sum_{H} \mathbf{P}(H \mid C)g_3(\text{:-}), H, C)$$

$$g_4(\text{:-}), C) = \sum_{H} \mathbf{P}(H \mid C)g_3(\text{:-}), H, C)$$

Last table to compute



The Function g_4

$$g_4(\text{:-}), C) = \sum_{H} \mathbf{P}(H \mid C)g_3(\text{:-}), H, C)$$

g_3(:-),H,C)			
:-)=t	H=p	C=t	0.5723
:-)=t	H=p	C=f	0.522
:-)=t	H=f	C=t	0.2468
:-)=t	H=f	C=f	0.2063
:-)=f	H=p	C=t	0.4278
:-)=f	H=p	C=f	0.478
:-)=f	H=f	C=t	0.752
:-)=f	H=f	C=f	0.793

g_4(:-),C)		
:-)=t	C=t	0.7(0.5723)+0.3(0.2468)=0.4747
:-)=t	C=f	0.4(0.522)+0.6(0.2063)=0.3326
:-)=f	C=t	0.7(0.4278)+0.3(0.752)=0.5251
:-)=f	C=f	0.4(0.478)+0.6(0.793)=0.667

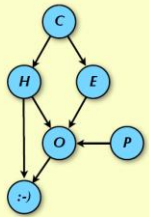
P(H|C)	H=p	H=f
C=t	0.7	0.3
C=f	0.4	0.6

Happiness and Class Attendance

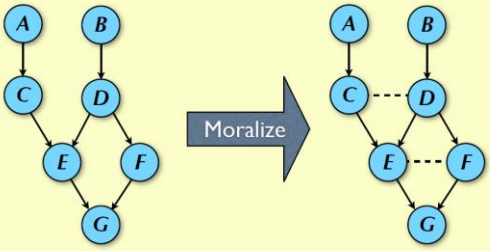
$$\mathbf{P}(:\text{-}), C) = Pr(C)g_4(:\text{-}), C)$$

$$\begin{aligned}\mathbf{P}(:\text{-}) \mid C) &= \frac{Pr(C)g_4(:\text{-}), C)}{\sum_{:\text{-})} Pr(C)g_4(:\text{-}), C)} \\ &= \frac{Pr(C)g_4(:\text{-}), C)}{Pr(C)\sum_{:\text{-})} g_4(:\text{-}), C)} \\ &= \frac{Pr(C)g_4(:\text{-}), C)}{Pr(C)} \\ &= g_4(:\text{-}), C)\end{aligned}$$

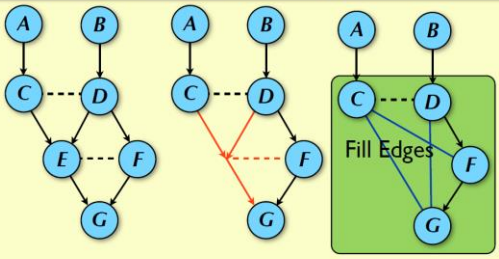
$g_4(:\text{-}), C)$		
:-)=t	C=t	0.4747
:-)=t	C=f	0.3326
:-)=f	C=t	0.5251
:-)=f	C=f	0.667



Variable Elimination as Pictures



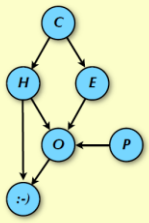
Variable Elimination and Fill Edges



$$g(C, D, G, F) = \sum_e P(E = e|C, D)P(G|E = e, F)$$

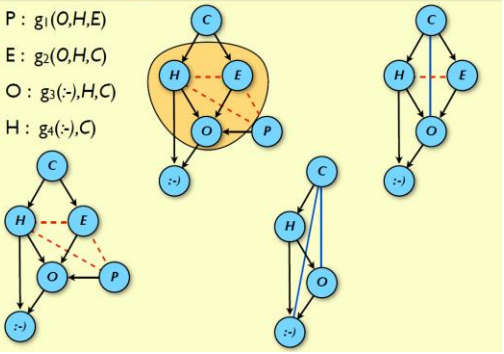
Recall The Happiness Problem

- We eliminated:
  - P,E,O,H
- Producing Functions
  - $g_1(O,H,E)$
  - $g_2(O,H,C)$
  - $g_3(:\text{-}),H,C)$
  - $g_4(:\text{-}),C)$

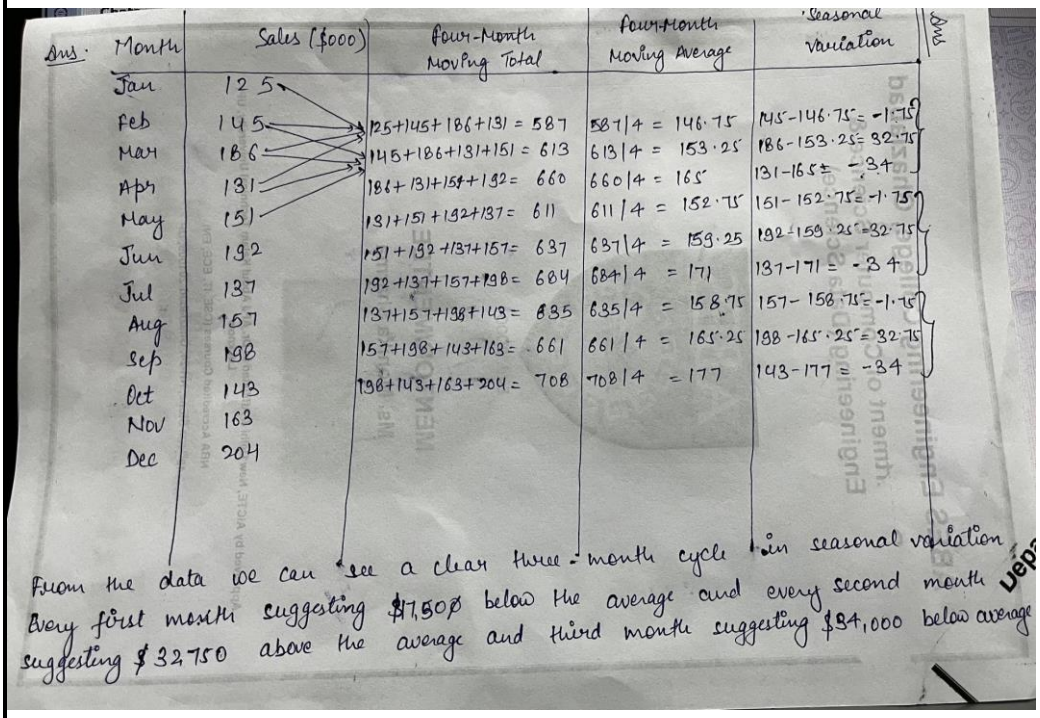


Recall The Happiness Problem

- P : $g_1(O,H,E)$
- E : $g_2(O,H,C)$
- O : $g_3(:\text{-}),H,C)$
- H : $g_4(:\text{-}),C)$



⇨ g1*g2*g3*g4 = 0.845*0.7445*0.5723*0.4747 = 0.1709

b)	Monthly sales revenue data were collected for a company: Calculate 4MA and trend with seasonal variation for given sales data. **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** Sales $000: 125 | 145 | 186 | 131 | 151 | 192 | 137 | 157 | 198 | 143 | 163 | 204  Ans. Month: Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec Sales ($000): 125, 145, 186, 131, 151, 192, 137, 157, 198, 143, 163, 204 Four-Month Moving Total: 125+145+186+131 = 587; 145+186+131+151 = 613; 186+131+151+192 = 660; 131+151+192+137 = 611; 151+192+137+157 = 637; 192+137+157+198 = 684; 137+157+198+143 = 635; 157+198+143+163 = 661; 198+143+163+204 = 708 Four-Month Moving Average: 587/4 = 146.75; 613/4 = 153.25; 660/4 = 165; 611/4 = 152.75; 637/4 = 159.25; 684/4 = 171; 635/4 = 158.75; 661/4 = 165.25; 708/4 = 177 Seasonal Variation: 145 − 146.75 = −1.75; 186 − 153.25 = 32.75; 131 − 165 = −34; 151 − 152.75 = −1.75; 192 − 159.25 = 32.75; 137 − 171 = −34; 157 − 158.75 = −1.75; 198 − 165.25 = 32.75; 143 − 177 = −34 From the data we can see a clear three-month cycle in seasonal variation. Every first month suggesting $1750 below the average and every second month suggesting $32,750 above the average and third month suggesting $34,000 below average **Note:** The four-month moving average represents the **trend.** We can see a clear trend in that every first month moving average is $6500 higher than the preceding month moving average, every second month moving average is $11750 higher than the preceding month moving average and every third month moving average is $12250 lower than the preceding month moving average. This trend can now be used to predict future underlying sales values. Once a trend has been established, any **seasonal variation** can be calculated. The seasonal variation can be assumed to be the difference between the actual sales and the trend (four-month moving average) value. Seasonal variations can be calculated using the additive or multiplicative models. **Using the additive model:** From the data we can see a clear three-month cycle in the seasonal variation. Every first month has a variation of -1.75, suggesting that this month is usually $1750 below the average. Every second month has a variation of 32.75 suggesting that this month is usually $32,750 above the average. In month 3, the variation of -34 suggests that every third month, the actual will be $34,000 below the average.	**5**	**CO2**	K 3	1.3.1

CO Course Outcomes mapped with respective question
KL Bloom's knowledge Level (K1, K2, K3, K4, K5, K6)
K1- Remember, K2- Understand, K3-Apply, K4- Analyze, K5: Evaluate, K6- Create