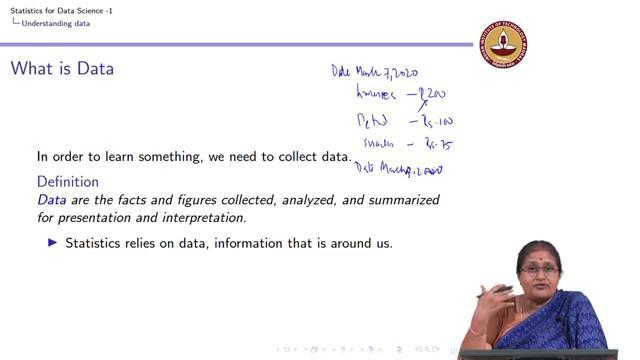
**स्टेटिस्टिक्स फॉर डाटा साइंस- 1 प्रो. उषा मोहन प्रबंधन अध्ययन विभाग भारतीय प्रौद्योगिकी संस्थान, मद्रास**

**लेक्चर - ०४**

**इंट्रोडक्शन एंड टाइप्स ऑफ़ डाटा पार्ट – २**

**इस अगले भाग में हम डाटा को समझने जा रहे हैं। यह हमारे लिए अत्यंत महत्वपूर्ण है क्योंकि स्टेटिस्टिक्स, डाटा पर निर्भर करती हैं और जब हम डाटा की बात करते हैं तो हम इन्फॉर्मेशन या सूचना की बात करते हैं जो हमारे चारों ओर मौजूद होती है। चाहे हम औपचारिक रूप से सांख्यिकीय विश्लेषण (statistical analysis) कर रहे हों या नहीं; हम सभी या तो डाटा बना रहे हैं या डाटा के संग्रह में योगदान दे रहे हैं या हम स्वयं डाटा एकत्र कर रहे हैं। उदहारण के तौर पर अगर हम साधारण घरेलू खातों के डाटा को ही ले जिसे हम हर दिन बनाते हैं।**

**(स्लाइड समय देखें: 00:52)**



**आमतौर पर हम तय करते हैं कि हमे हर दिन कितना खर्च करना हैं, हमारे आस-पास ऐसे लोग हैं जो खाते का हिसाब रखते है और खाते इस तरह के हो सकते हैं जैसे कि हम हर दिन कितना कर्च करते हैं। मानिए जैसे कि हम एक तारीख लिख लेते हैं और उसके बाद हम किराने का सामान 200 रुपये, फिर पेट्रोल 100 रुपये आदि लिखते हैं।**

**यहाँ हम एक तारीख लिख देते हैं, जैसे कि हमने यहाँ एक तारीख डाल दी है मार्च-सात-2020 और इससे बाद जैसे हमने स्नैक्स पर कुछ खर्च किया होगा; तो हम यहाँ स्नैक्स पर पिछत्तर (75) रुपये खर्च किए हैं लिख देते हैं। फिर से वापस आ कर 2 दिन बाद यानि 9-मार्च-2020 लिख देते हैं। हम फिर से अपने डाटा की गणना करते हैं, गणना के बाद मेरे पास फिर से डाटा है; यह सब भी एक डाटा है। हम सभी हर समय डाटा एकत्र करते हैं। हमारे आस-पास बहुत सारा डाटा है और हम सब डाटा में योगदान भी दे रहे हैं, हर बार जब हम कीबोर्ड का बटन या माउस पर क्लिक करते हैं, तो हम कुछ डाटा जेनेरेट कर रहे होते हैं।**

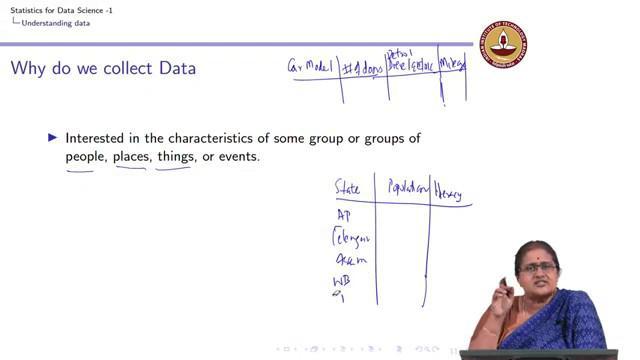
**जैसा कि मैंने पहले कहा कि पिछले कुछ वर्षों में स्टेटिस्टिक्स की परिभाषा में काफी बदलाव आया है। तो इन वर्षों मे डाटा की प्रकृति मे भी बदलाव हुआ है। लगभग पचास (50) वर्ष पहले जिसे हम डाटा के रूप मे जानते थे वह सिर्फ संख्या और कटेगोरिकल डाटा होता था। आज लोग सोशल मीडिया एनालिटिक्स, मल्टीमीडिया एनालिटिक्स, टेक्स्ट एनालिटिक्स के बारे में बात करते हैं और हर पल इतना डाटा जेनेरेट होता है कि यहां तक कि यूट्यूब वीडियो या मल्टीमीडिया वीडियो या फोटो पर टिप्पणी भी डाटा है। आप किसी उत्पाद ई-कॉमर्स पोर्टल पर आने वाली टिप्पणियों को देखते हैं जो डाटा ही है।**

**तो जो भी हम बोलते हैं और जब भी हम बोलते हैं वो भी डाटा के रूप मे एकत्र किया जा रहा है। हमारे आस-पास इतना अधिक डाटा है जो दिन प्रतिदिन उत्पन्न हो रहा है, इतना डाटा है जिसे एकत्र किया जा रहा है और इतना डाटा है जो सार्थक जानकारी में संसाधित (process) होने की प्रतीक्षा कर रहा है। यही स्टेटिस्टिक्स का उद्देश्य है और जब भी हम एक सांख्यिकीय विश्लेषण करना चाहते हैं तो हम डाटा पर बहुत अधिक निर्भर करते हैं।**

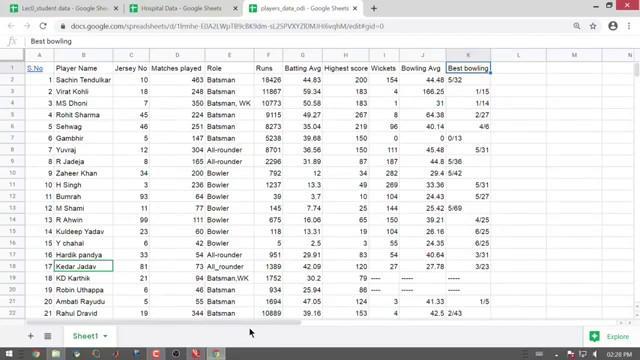
**तो आइए सबसे पहले हम डाटा को परिभाषित करते हैं कि डाटा क्या है, बहुत ही सरल शब्दों में कहें तो डाटा केवल तथ्य और एकत्र किए गए आंकड़े हैं; तथ्यों से हमारा मतलब है कि यह संख्यात्मक हो सकता है, यह किसी भी प्रकार का हो सकता है। इसलिए, जैसे कि मैंने अभी कहा कि मल्टीमीडिया या वीडियो या इंटरनेट या किसी उत्पाद पर या किसी भी चीज़ पर की गई टिप्पणियां आदि डाटा ही है।**

**तो, डाटा क्या है? डाटा तथ्य है; जो यथार्थ मे मौजूद है। हम इस डाटा को संक्षेप में प्रस्तुत करना चाहते हैं, हम प्रस्तुति और व्याख्या के लिए डाटा का विश्लेषण करना चाहते हैं। यह कहना बहुत गलत है कि हम चाहते हैं कि हमारा डाटा कुछ बताए। हम नहीं चाहते हैं कि हमारा डाटा कुछ कहे, हम वास्तव में, डाटा को एक तथ्य के रूप मे देखते हैं जिसका उपयोग हम वो जानकारी निकालने के लिए करते हैं जो डाटा व्याख्या के उद्देश्यों के लिए कह रहा है। यह वह ज्ञान है जो जिसे हमें समझने की आवश्यकता है कि डाटा क्या है और इस मॉड्यूल में, हम यही समझने जा रहे हैं कि डाटा क्या है।**

**(स्लाइड समय देखें: 03:59)**



**(स्लाइड समय देखें: 04:08)**



**हम डाटा क्यों एकत्र करते हैं? जब हम अपने क्रिकेटिंग डाटा सेट पर वापस जाते हैं तो हम जानना चाहते कि यह डाटा क्यों एकत्र किया गया है? इस डाटा को एकत्र करने का उद्देश्य क्या है? इस डाटा को देखते हुए हम कुछ सवालों के जवाब देना चाहेंगे जैसे कि वह व्यक्ति कौन है जिसने सबसे ज्यादा मैच खेले हैं। यह एक बहुत ही स्वाभाविक सवाल है, या जैसे कि किस व्यक्ति के पास सबसे ज्यादा बल्लेबाजी औसत है या हम यह जानना चाहें कि सबसे ज्यादा विकेट लेने वाला व्यक्ति कौन है?**

**तो आप देख सकते हैं कि हम यहां निश्चित रूप से यह सब जानना चाहेंगे जैसे कि हम यहाँ देख सकते हैं कि सचिन तेंदुलकर ने चार सौ तिरेसठ (463) मैच खेले हैं, लेकिन फिर हम हर उस मैच के बारे में भी डाटा लेना चाहेंगे जो तेंदुलकर ने इन 463 मैचों में खेला है। हो सकता है हम यह देखना चाहे कि उनका बल्लेबाजी प्रदर्शन हर मैच मे कैसा है या उनका ओवरऑल प्रदर्शन क्या है और उन्होंने वह मैच कहाँ खेले है, यह देखने के लिए कि क्या उनके घरेलू बल्लेबाजी का प्रदर्शन देश के बाहर बल्लेबाजी के प्रदर्शन से भिन्न है। शायद हम हर खिलाड़ी के साथ ऐसा डाटा जेनेरेट करना चाहे। तो आप देख सकते हैं कि यहाँ हमारे सामने बहुत सारे प्रश्न उत्पन्न हो रहे हैं और इन सभी प्रश्नों के उत्तर के लिए हमें डाटा की आवश्यकता होगी।**

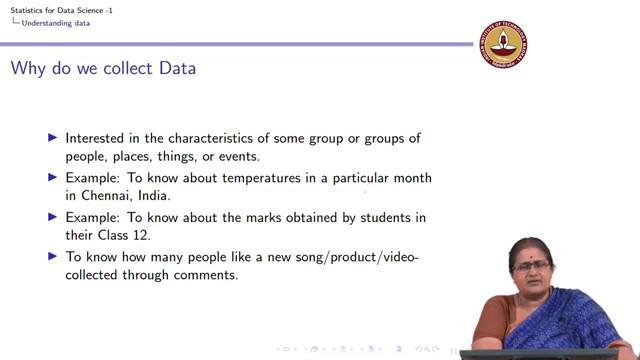
**तो अगर हम फिर से एक प्रश्न करे कि हम डाटा क्यों एकत्र करते हैं? हमारे डाटा एकत्र करने का प्राथमिक कारण है कि हम समूहों की विशेषताओं के बारे में जानने में रुचि रखते हैं; यह लोगों के समूह हो सकते हैं, यह स्थान के समूह हो सकते हैं, यह चीजें हो सकती हैं, यह घटनाएं भी हो सकती हैं। यहाँ ध्यान दें कि हम हमेशा सिर्फ लोगों के बारे मे जानने में रुचि नहीं रखते हैं।**

**उदाहरण के लिए हमारे पास डाटा का एक संग्रह हो सकता है जिसमें हमारे पास कार का मॉडल हो, उस कार के दरवाजों की संख्या हो, चाहे वह कार डीजल हो, चाहे वह पेट्रोल हो, चाहे वह एक इलेक्ट्रिक कार हो। तो, यहाँ आप देखते हैं कि यहाँ हमारे पास इस डाटा मे कोई भी लोग शामिल नहीं हैं या लोगों का समूह शामिल नहीं है, इसके अलावा और भी प्रश्न हो सकते हैं जैसे कि कार का माइलेज कितना है, क्या कार सेडान है या हैचबैक; ये सभी कुछ ऐसी चीजें हैं जिन्हें हम यहाँ इकट्ठा करने जा रहा हैं।**

**तो आप यहाँ देखें कि जब भी हम लोगों के समूह के बारे मे बात करते हैं तो यह चीजें भी हो सकती हैं, एक डाटा हो सकता है जिसे हम एकत्र कर रहा हो जिसमें राज्य के नाम और उनकी जनसंख्या शामिल हो, उनकी साक्षरता दर शामिल हो जैसे राज्य के नाम आंध्र प्रदेश, तेलंगाना, असम, पश्चिम बंगाल, तमिलनाडु हो सकते हैं; हम यहाँ जनसंख्या को नोट कर सकते हैं और साक्षरता दर को भी।**

**यहां आप देख सकते हैं कि हमे राज्यों के बारे में जानने में दिलचस्पी है, इसलिए डाटा स्थान हो सकते हैं; यह कुछ भी हो सकते है। इसलिए, जब हम डाटा को एकत्र करने मे खुद को सीमित कर रहे हैं तो यह केवल लोगों तक ही सीमित नहीं है। तो हम वही प्रश्न दोबारा पूछते हैं हम डाटा क्यों एकत्र करते हैं? जब भी हम किसी विशेषता (characteristics) या गुण (attributes) में रुचि रखते हैं तो हम डाटा एकत्र करते हैं और हम इस डाटा से इन विशेषताओं या गुणो के बारे में उत्तर पता करने मे रुचि रखते हैं।**

**(स्लाइड समय देखें: 07:30)**

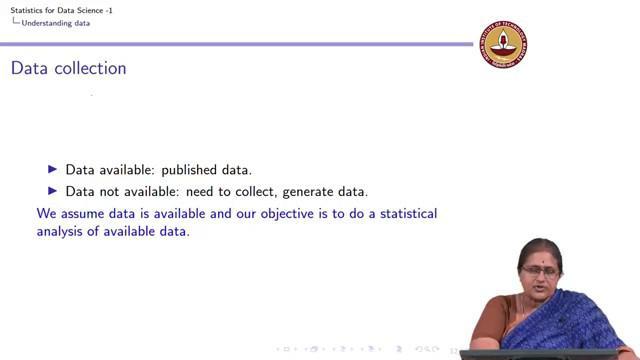


**उदाहरण के लिए जैसे हम चेन्नई में किसी विशेष महीने में तापमान के बारे में जानना चाहते हैं तो चेन्नई फिर से एक जगह है और यहां हमे लोगों के बारे में कोई दिलचस्पी नहीं है। हम छात्रों के द्वारा प्राप्त अंकों के बारे में जानना चाहे या यह जानना चाहे कि कितने लोग एक नया गीत पसंद करते हैं। इन दिनों इंटरनेट पर स्ट्रीमिंग कर रहे कंटैंट पर आपको पसंद- नापसंद का डाटा मिल जाएगा। तो आप जानना चाह सकते हैं कि कितने लोग एक नया गाना, नया उत्पाद, नया वीडियो पसंद करते हैं। इसे टिप्पणियों के माध्यम से पूरी तरह से अलग तरीके से एकत्र किया जाता है।**

**तो जब हम तापमान के बारे में बात करते हैं तो आप देख सकते हैं कि इस डाटा को एकत्र करने का एक तरीका है। जब हम छात्रों के अंकों के बारे में बात करते हैं तो इस तरह के डाटा को एकत्र करने का एक तरीका होता है, अंक यहाँ या तो प्रतिशत हो सकते हैं या यह ग्रेड हो सकते हैं। जब हम अंकों की बात करते हैं तो हमेशा अंकों प्रतिशत मे नहीं होते है और आजकल तो अधिकांश बोर्ड और कॉलेज प्रतिशत अंक नहीं देते हैं। ऐसे बहुत से लोग हैं जिन्होंने छात्रों का मूल्यांकन करने के लिए ग्रेड या लेटर ग्रेड पर स्विच किया है।**

**तो आप देख सकते हैं कि यह सब डाटा का गठन करता है; टिप्पणियाँ भी डाटा है, लेकिन यह डाटा का एक बिल्कुल अलग रूप है, यह टेक्स्ट डाटा है। इसलिए जैसे ही हम डाटा के बारे में बात करते हैं, हम देखते हैं कि डाटा हमारे चारों ओर है। हमें यह समझने की आवश्यकता है कि हम डाटा एकत्र कर रहे हैं क्योंकि हम वास्तव मे समूहों या लोगों या घटनाओं की विशेषताओं में रुचि रखते हैं। इसलिए, हम डाटा एकत्र करते हैं।**

**(स्लाइड समय देखें: 09:26)**



**अगली महत्वपूर्ण बात यह है कि हम यह डाटा कहां से एकत्र करेंगे? हमे यह डाटा कहां से मिलेगा? इस प्रश्न का उत्तर देने के लिए आप देख सकते हैं कि या तो आप जाकर डाटा एकत्र करें या ये डाटा एकत्र करने के लिए पहले से ही जो डाटा उपलब्ध है यानि जो प्रकाशित डाटा है उसे उपयोग करे। प्रकाशित डाटा के लिए आप सरकारी साइट पर जा सकते हैं। अधिकांश सरकारें अपनी डाटा साइट प्रकाशित करती हैं।**

**(स्लाइड समय देखें: 09:57)**

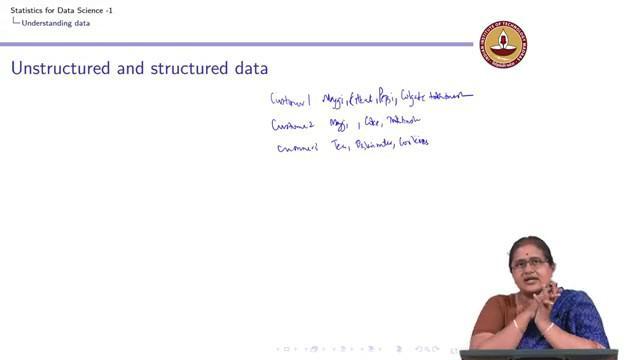


**अगर आप यहाँ देखे तो यह एक साइट data.gov है। यह भारत का एक खुला सरकारी डाटा प्लेटफॉर्म है। यह डाटा साइट लगभग वो सभी डाटा देती है जो सरकारी स्तर पर एकत्र किया जाता है। आप देख सकते हैं कि आपके पास यहाँ पीने का पानी और स्वच्छता का डाटा, स्वास्थ्य डाटा, आपके पास इकॉनमी का डाटा, परिवहन डाटा, शिक्षा का डाटा है; आपके पास यहां काफी डाटा उपलब्ध है।**

**तो यहां मुख्य विचार जो मैं आपको बताना चाहती हूं वह यह है कि या तो आप जाकर डाटा एकत्र करे या फिर प्रकाशित डाटा का उपयोग कर डाटा पर काम कर सकते हैं। यदि आप जिन प्रश्नों का उत्तर देना चाहते हैं, उन्हें डाटा की आवश्यकता है जिसे उत्पन्न या एकत्र किया जाना है तो आपको जाकर डाटा एकत्र करना होगा।**

**यह पाठ्यक्रम आपको यह समझने के लिए प्रेरित नहीं करेगा कि डाटा को कैसे एकत्र किया जाए या यह डाटा कैसे उत्पन्न किया जाए, क्योंकि हम यहाँ यह मान के चल रहे हैं कि डाटा हमारे लिए उपलब्ध है और हमारा उद्देश्य उपलब्ध डाटा का सांख्यिकीय विश्लेषण या स्टेटिस्टिक्ल एनालिसिस करना है। इस पाठ्यक्रम** का **यही उद्देश्य है। फिर भी यदि आप उन प्रश्नों के उत्तर मांग रहे हैं जिनके लिए डाटा उपलब्ध नहीं है, तो आपको यह समझने की आवश्यकता है कि इस डाटा को संरचित या वैज्ञानिक तरीके से कैसे एकत्रित किया जाए, कैसे डाटा को उत्पन्न किया जाए। यह भी आपको आने वाले समय में सिखाया जाएगा।**

**(स्लाइड समय देखें: 11:41)**

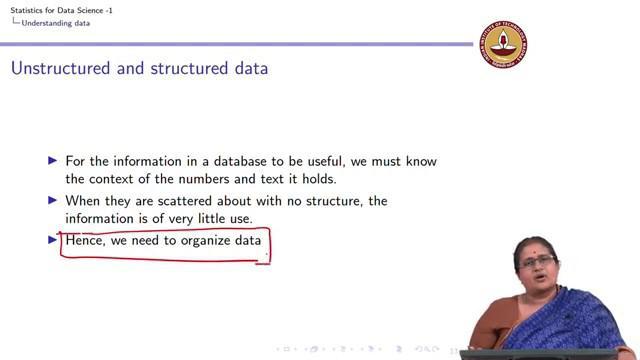


**मान लीजिए मेरे पास एक डाटा कि फाइल है जो इस तरह की है, मानिए कि हम एक खुदरा बाजार (retail market) में एक व्यक्ति से पूछते हैं कि कौन सी चीजें बेची गई हैं। और वह व्यक्ति इस तरह के डाटा को आपके सामने प्रस्तुत करता है की ग्राहक एक (1) ने मैगी, किटकैट, पेप्सी और शायद कुछ कोलगेट टूथब्रश खरीदा। ग्राहक दो (2) ने मैगी, कोक, टूथब्रश खरीदा। ग्राहक तीन (3) ने चाय खरीदी, फिर बिसलेरी का पानी और कुछ कुकीज खरीदीं।**

**तो यह वह डाटा है जो एक खुदरा काउंटर पर बैठा व्यक्ति एकत्र कर रहा है और यदि वह आपके सामने इस तरह का कुछ डाटा प्रस्तुत करता है तो आप इस डाटा से कोई मतलब नहीं निकाल सकते। इस बात का मतलब यह है की यह आपके लिए डाटा ही है लेकिन क्या हम इस डाटा का कोई अर्थ निकाल सकते हैं? मान लीजिए कि वह व्यक्ति जिसने यह ग्राहक डाटा एकत्र किया है, वह हमें पचास (50) ग्राहकों के लिए समान डाटा दे रहा है। ठीक है?**

**तो तुरंत यह डाटा कहलाएगा, लेकिन यह एक असंरचित रूप में है। यह आंकड़े असंगठित रूप में है। यह डाटा ही कहलाएगा लेकिन यह बहुत संरचित या संगठित रूप में नहीं है। इस पाठ्यक्रम में हम केवल उन आंकड़ों का विश्लेषण करना चाहेंगे जो हमारे पास संरचित रूप में आते हैं।**

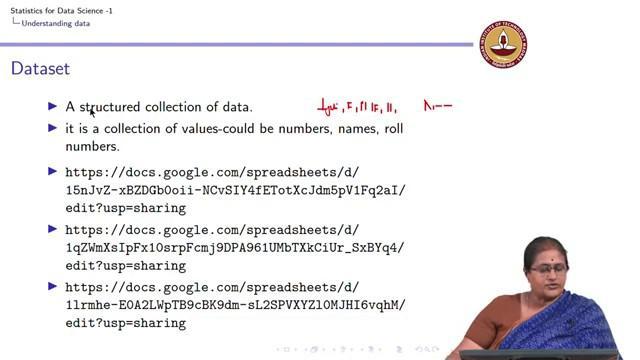
**(स्लाइड समय देखें: 13:48)**



**इसलिए किसी डाटा की जानकारी के उपयोगी होने के लिए हमें उस डाटा मे संख्याओं और टेक्स्ट के संदर्भ (context) को जानना चाहिए। जब डाटा असंचरित या अनऑर्गनाइज़ हो जैसा की मैंने अपने पहले उदाहरण मे बताया, बिना किसी संरचना के जानकारी बहुत कम उपयोग की है; हालांकि, हमे डाटा को व्यवस्थित करने की आवश्यकता है। तो अब हमारे लिए अनिवार्य आवश्यकता डाटा को व्यवस्थित करने की है। यह सबसे महत्वपूर्ण चीज है जो हमे करने की जरूरत है।**

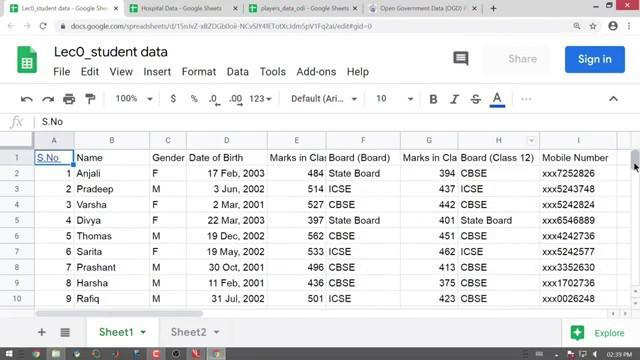
**डाटा व्यवस्थित करने से मेरा क्या मतलब है?**

**(स्लाइड समय देखें: 14:25)**



**आइए यहाँ हम डाटा के संरचित (ऑर्गनाइज़ड) संग्रह को देखते हैं।**

**(स्लाइड समय देखें: 14:38)**

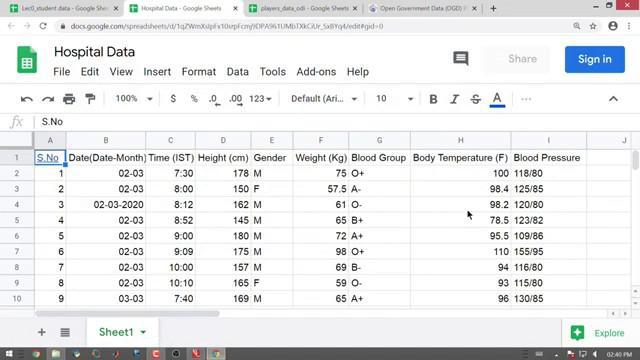


**यदि आप डाटा के एक संरचित संग्रह को ले, जैसा की आप यहाँ देख सकते हैं कि यह वही डाटा सेट है जिसे हमने पिछली बार जब पाठ्यक्रम शुरू किया था तो देखा था। आप पहले से ही जानते हैं की यह एक काल्पनिक डाटा सेट है। कल्पना कीजिए कि अगर यह डाटा सेट एक व्यक्ति द्वारा एकत्र किया गया था जब प्रत्येक छात्र कॉलेज में प्रवेश कर रहा था। जैसे कि किसी ने कॉलेज में प्रवेश किया तो उस व्यक्ति ने लिखा - अंजलि, महिला, बोर्ड आईसीएसई, प्राप्त अंक 98 और फिर रामू, पुरुष, इत्यादि तो अगर उस व्यक्ति ने इस तरह से यह डाटा संग्रह किया होता तो यह डाटा फिर से असंरचित (अनऑर्गनाइज़ड) हो जाता।**

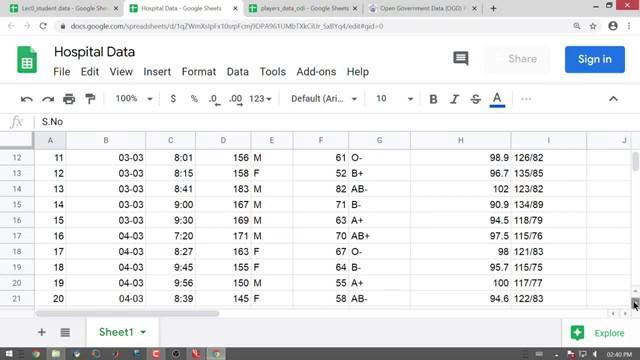
**तो हम यहाँ इस डाटा को एक डाटा तालिका के रूप में संदर्भित करने के लिए, एक संरचना देने की कोशिश कर रहे हैं और इस पाठ्यक्रम में हम अपने डाटा का विश्लेषण करने के लिए बड़े पैमाने पर Google शीट का उपयोग करेंगे। तो हम इस डाटा को तालिका मे प्रस्तुत करना चाहते हैं। हमारे पास डाटा को सारणीबद्ध (tabulate) करने के कई तरीके हैं।**

**अब हम यहाँ आपको यह बताने जा रहे हैं कि इस पाठ्यक्रम में डाटा का विश्लेषण करने के लिए हम किस तरीके का उपयोग करेंगे और मुख्य रूप से हम डाटा का विश्लेषण करने के लिए यहाँ Google शीट का उपयोग करेंगे। तो अब हम यह समझने जा रहे हैं कि हम स्प्रेडशीट में डाटा को कैसे व्यवस्थित करेंगे। यही हमारा अगला कदम होगा।**

**(स्लाइड समय देखें: 16:24)**

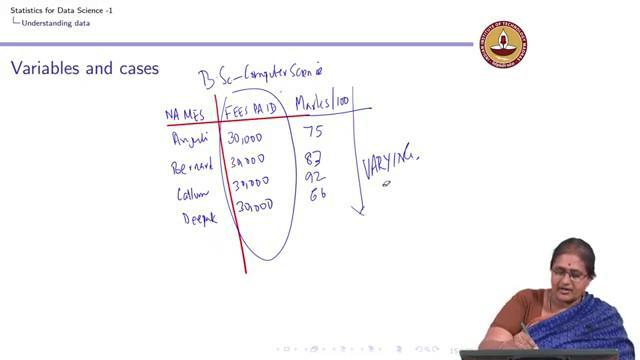


**(स्लाइड समय देखें: 16:32)**



**आप यदि यहां एक अन्य डाटा सेट को देखें तो आप पाएंगे कि इस डाटा सेट मे लगभग नौ (9) रोगियों का डाटा एकत्र किया गया है; क्षमा करें, इसने लगभग बीस (२०) रोगियों का डाटा एकत्र किया है जो ७:३०, ८:००, ८:१२ बजे एक डायग्नोस्टिक केंद्र मे प्रवेश करते हैं; रोगियों की ऊंचाई, जेंडर, रक्त समूह (blood group), शरीर का तापमान और रक्तचाप आदि डाटा एकत्र किया गया है। तो, आप फिर से यहाँ पाते हैं कि यह हमें एक प्रकार का सारणीबद्ध डाटा दिया गया है। अब हम यह समझने जा रहे हैं कि डाटा के संरचित (structured) संग्रह से हमारा क्या मतलब है और अब हम इसी पर अपना ध्यान केंद्रित करेंगे।**

**(स्लाइड समय देखें: 17:11)**



**जब हम डाटा सेट बनाने के लिए डाटा के संरचित संग्रह की बात करते है, तो हमे सबसे पहले यह समझने की जरूरत है कि चर यानि वेरिएबल क्या है।**

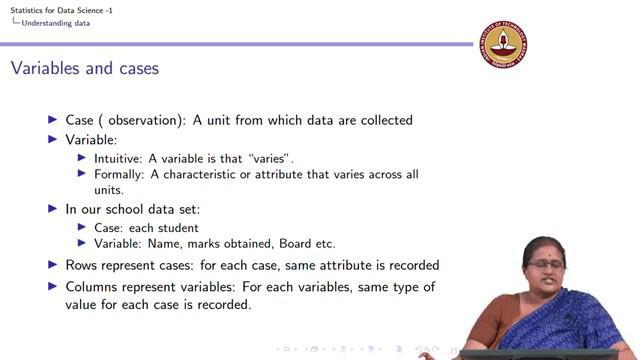
**मान लीजिए कि मेरे पास फिर से डाटा का एक सेट है जो इस तरह दिखता है यहाँ नाम है और हम यहाँ भुगतान की गई फीस को देख रहे हैं। कोर्स मान लीजिए कि बीएससी कंप्यूटर साइंस कोर्स है। यह फिर से काल्पनिक डाटा है। मेरे पास नाम हैं जैसे अंजलि, बर्नार्ड, कैलम, दीपक वगैरह।**

**आप भुगतान की गई फीस को देखे, फिर से ये काल्पनिक संख्याएं ही हैं। अगर मैं सभी की भुगतान की गयी फीस देखूँ तो यह तीस हज़ार रुपये (३०००० INR)** है। **सभी की तीस तीस तीस हज़ार (३००००, ३००००, ३००००) मैं देख रही हूँ; और अगली चीज जो मैं देखने जा रही हूं, वह यह है कि 100 में से प्राप्त अंक क्या हैं, तो मेरे पास अंक 75, 83, 92, 66 हो सकते हैं।**

**जब हम इस तालिका (table) को देखते हैं तो आप पाते हैं कि वे सभी छात्र समान शुल्क का भुगतान कर रहे हैं। इसलिए भुगतान की गई फीस के संबंध में हम कुछ भी और नहीं पूछना चाहते। हर कोई एक ही फीस दे रहा है, कुछ भी नहीं बदल रहा है यह इन सभी लोगों के साथ स्थिर है। लेकिन जब हम उनके अंकों को देखते हैं तो पाते हैं कि अंजलि ने 75 अंक प्राप्त किए हैं, बर्नार्ड ने 83 अंक प्राप्त किए हैं, कैलम ने 92 अंक प्राप्त किए हैं और दीपक ने 66 अंक प्राप्त किए हैं; दूसरे शब्दों में यह भिन्न है या वेरी (vary) कर रहा है।**

**दूसरे शब्दों मे, हमारे पास वहां परिवर्तनशीलता की अवधारणा (concept of variability) है। तो जब हम पूछते हैं कि चर यानि वेरिएबल क्या है तो उत्तर यहाँ फिर से बहुत सरल है, हम यहाँ केवल एक बुनियादी स्तर पर परिचय दे रहे हैं कि चर यानि वेरिएबल क्या है। तो आइये हम यहाँ चर को निम्नलिखित रूप में परिभाषित करते हैं।**

**(स्लाइड समय देखें: 19:29)**



**वेरिएबल यानि चर को "भिन्न" के रूप में परिभाषित किया जा सकता है जो बदलती रहती है और औपचारिक रूप से यह एक विशेषता या गुण है जो सभी इकाइयों में भिन्न होती है।**

**आइए अब समझते हैं कि मेरे द्वारा अब तक वर्णित प्रत्येक डाटा सेट में एक इकाई क्या है और चर क्या है। अब यदि हम इस डाटा सेट तो देखे तो आप पाते हैं कि नाम यहाँ सही मायने में चर है, जेंडर निश्चित रूप से वेरिएबल है। यदि हम यह डाटा केवल पुरुषों के कॉलेज या केवल लड़कियों के स्कूल से ले रहे होते तो जेंडर वेरिएबल नहीं होता; वो सब इकाइयों यानि यूनिट के लिए एक ही होता। लेकिन यहां इस उदाहरण मे आप देख सकते हैं कि यह परिवर्तनशील है, प्राप्त अंक फिर से परिवर्तनशील है।**

**क्या जन्म की तारीख हर किसी व्यक्ति की समान होगी, हर व्यक्ति का जन्म एक ही दिन नहीं हुआ होगा यह बहुत संभव हो सकता है यदि जन्म का वर्ष लिया गया हो तो शायद यह बहुत भिन्न न हो, लेकिन फिर निश्चित रूप से वहां कुछ हल्की मात्रा में परिवर्तनशीलता होगी; फिर बोर्ड, मोबाइल नंबर आदि आएंगे।**

**तो यहाँ यदि आप इसे इस तरह से देखें कि हमने नाम, जेंडर, जन्म तिथि, अंक, बोर्ड को परिभाषित किया है और सभी वेरिएबल हैं, जबकि यहाँ अंजलि, प्रदीप, वर्षा, दिव्या सभी केसेस या अवलोकन (observation) हैं। यहाँ प्रत्येक केस के लिए हमने हर वेरिएबल को दर्ज़ किया है जैसे अंजलि के लिए हमने प्रत्येक वेरिएबल दर्ज किया है।**

**इसी तरह हम अब अस्पताल के आंकड़ों पर आते हैं। यहाँ चर समय है, तारीख है; यदि आप उस तिथि को देखेंगे तो आप पाएंगे की तारीख एक ही है। तो यहाँ हम देख सकते हैं कि तारीख अलग-अलग नहीं है, यहाँ यह तीसरी है। यदि मेरा डाटा एक सबसेट है और हम केवल पहले आठ अवलोकनों (observation) को देखे तो यह उसी एक तारीख को लिए गए हैं। तो तारीख बदल नहीं रही है, जबकि प्रविष्टियों (entries) का समय अलग-अलग है। यह भी डाटा ही है; ऊंचाई भिन्न (vary) होती है, जेंडर भिन्न होता है, रक्त समूह भी भिन्न भिन्न होता है, इसी तरह रक्तचाप और शरीर का तापमान भी भिन्न होता है।**

**आप देख सकते हैं कि इस केस में मैंने किसी व्यक्ति का नंबर नोट नहीं किया है, लेकिन हम उन्हें उनके आगमन पर व्यक्ति एक (1) कह कर संबोधित कर सकते हैं। उनकी ऊंचाई, जेंडर, वजन, रक्त समूह, शरीर का तापमान और रक्तचाप उनके आगमन के समय दर्ज किया गया है। इसी तरह प्रत्येक व्यक्ति जो सिस्टम में प्रवेश करता है उस व्यक्ति का प्रत्येक वेरिएबल दर्ज किया जाता है।**

**इसी तरह प्लेयर्स डाटा सेट में हमारे पास जर्सी नंबर है। अब दिलचस्प बात यह है कि किसी भी दो खिलाड़ियों के पास एक ही जर्सी नंबर नहीं होता है और हर किसी के पास एक अलग जर्सी नंबर होता है, साथ ही मे खेले गए मैच, खिलाड़ी की भूमिका, देश का नाम भी यहां वेरिएबल हो सकता है, लेकिन चूंकि हमारे पास जो डाटा है वह केवल भारत का है तो देश यहाँ परिवर्तनशील नहीं है; इसी तरह उच्चतम स्कोर, विकेट, गेंदबाजी औसत वेरिएबल हो सकते हैं ।**

**अब एक और दिलचस्प बात यह है कि इस डाटा सेट में आपको कुछ इस तरह का डाटा मिलता है जहां आप पाते हैं कि एक डाटा है जो शून्य (0) बता रहा है और एक डाटा है जो सिर्फ डैश (---) दिखा रहा है। अब इसका क्या मतलब है? आपने देखा कि गौतम गंभीर ने कोई डाटा नहीं लिया है, लेकिन उन्होंने गेंदबाजी तो की है, ठीक है? तो, उनके लिए यहाँ मान शून्य (0) है, जबकि, जब आप दिनेश कार्तिक या रॉबिन उथप्पा के डाटा को देखते हैं, तो आप देखते हैं कि उनके पास वह डाटा नहीं है जो उन्हें कोई भी गेंदबाजी आँकड़े देने के लिए उपलब्ध हो।**

**इसी तरह किसी भी मामले में भले ही हम डाटा एकत्र कर रहे हो पर वहां यह संभव हो सकता है कि जिस डाटा को हम वेरिएबल के एक सबसेट के रूप मे तलाश कर रहे हैं वह हर इकाई के लिए उपलब्ध ही न हो , जैसा कि हमने इस मामले में देखा है, ठीक है?**

**तो, यह वही है जिसका हम उल्लेख कर रहे थे जहां डाटा उपलब्ध ही नहीं है। फिर भी ये लोग डाटा सेट का एक बहुत बड़ा हिस्सा हैं। उनका बल्लेबाजी औसत मौजूद है, लेकिन उनका गेंदबाजी औसत उपलब्ध नहीं है। इस बिंदु पर मैं जिस बात पर जोर देना चाहती हूं वह यह है कि डाटा की अनुपलब्धता शून्य (0) मान लेने वाले डाटा से अलग है; इसके बहुत मायने हैं। यहाँ शून्य (0) लिया गया है, भले ही एक व्यक्ति ने तेरह (13) ओवर फेंके हैं, जबकि दूसरी तरफ यहां कुछ खिलाड़ियों के आंकड़े उपलब्ध नहीं हैं क्योंकि इन खिलाड़ियों ने गेंदबाजी की ही नहीं है। हम इसे शून्य (0) के रूप में नहीं ले सकते हैं क्योंकि इससे पूरी बात ही अलग हो जाएगी।**

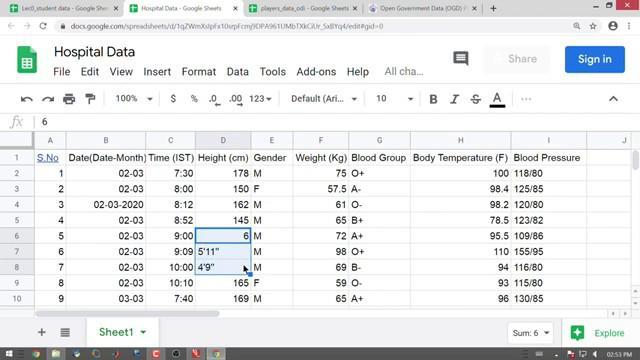
**तो इसलिए जब हम डाटा को देखते हैं तो सबसे पहले हमें यह समझने की आवश्यकता होती है कि अवलोकन क्या है और चर क्या है। सहज रूप से एक चर वह है जो औपचारिक रूप से भिन्न होता है यह एक विशेषता है जो सभी इकाइयों में भिन्न होती है। यदि उस इकाई के लिए विशेषता उपलब्ध है, जैसा कि हमने क्रिकेट डाटा सेट में देखा कि एक विशेषता जैसे की गेंदबाजी का औसत, कुछ खिलाड़ियों के लिए उपलब्ध ही नहीं है।**

**इसी तरह हमारे स्कूल डाटा सेट में प्रत्येक छात्र एक केस था जहां वेरिएबल नाम, प्राप्त किए गए अंक, बोर्ड का नाम वगैरह थे। पंक्तियाँ (rows) प्रत्येक केस का प्रतिनिधित्व करती हैं जिनमे हर एक केस की सभी विशेषता दर्ज की जाती है या हमे वह विशेषता उपलब्ध नहीं है। क्रिकेटिंग डाटा सेट में कुछ खिलाड़ियों के लिए गेंदबाजी तथा गेंदबाजी का औसत की विशेषता उपलब्ध नहीं थी। इसलिए हम यह उपलब्ध नहीं के रूप में रिकॉर्ड करते हैं यदि वह विशेषता उपलब्ध नहीं है और हम इसे शून्य (0) के रूप में रिकॉर्ड करते हैं यदि यह मान शून्य (0) है। तो मान शून्य (0) और उपलब्ध नहीं होने के बीच अंतर है।**

**कॉलम चर यानि वेरिएबल का प्रतिनिधित्व करते हैं और प्रत्येक वेरिएबल के लिए एक ही प्रकार का मान दर्ज किया जाता है। एक ही प्रकार के मान से हमारा क्या तात्पर्य है? इसको समझने के लिए आइये फिर से हम अपने अस्पताल के डाटा सेट पर वापस चलते हैं। इस अस्पताल डाटा** के **सेट में, यहाँ दो वेरिएबल हैं, एक है ऊँचाई (height) और दूसरा वज़न (weight) है।**

**ऊंचाई के आगे आपको सेंटीमीटर लिखा दिखाई दे रहा होगा इसी तरह वज़न के आगे आप देख सकते हैं कि किलोग्राम लिखा है। आप देख सकते हैं कि यह डाटा संग्रह 7:30 बजे शुरू होता है और 2 मार्च को 10 बजे समाप्त होता है। मान लीजिए यहाँ हमारे पास दो (2) घंटे की शिफ्ट में काम करने वाले लोग हैं, तो 7 बजे शिफ्ट शुरू करने वाला व्यक्ति 9 बजे अपनी शिफ्ट समाप्त करेगा और 9 बजे अपनी शिफ्ट शुरू करने वाला व्यक्ति 11 बजे तक काम करेगा। तो यदि जो व्यक्ति 9 बजे शिफ्ट में आता है वह आने वाले रोगियों की ऊंचाई को फीट में उल्लेख करना चाहे तो वह यहाँ 180 सेंटीमीटर के बजाय 6 फीट के रूप में उल्लेख करेगा।**

**(स्लाइड समय देखें: 27:25)**

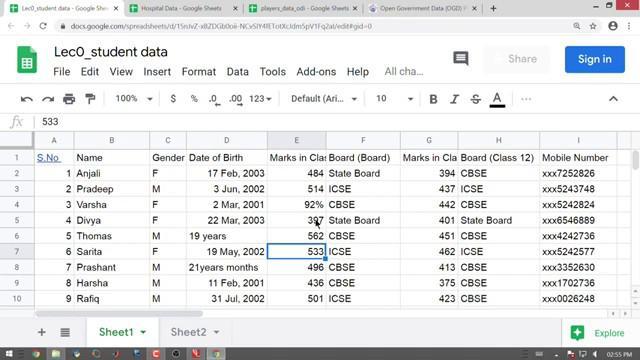


**तो अचानक आपका डाटा सेट एक निश्चित बिंदु के बाद छह (6) के रूप में दिखना शुरू हो जाएगा। आप डाटा सेट को देखगे तो पाएंगे यह 6 होगा; यह कोई पाँच फीट ग्यारह इंच (५ फुट ११ इंच) हो सकता है और यह चार फुट नौ इंच (४ फुट ९) इंच के करीब।**

**अब 10 बजे एक और व्यक्ति आता है और यही काम फिर से शुरू करता है। तो आप यहाँ देख सकते हैं कि यह तीन डाटा इकाइयाँ (unit) भले ही ऊँचाई माप रही हों लेकिन जिन इकाइयों का उपयोग किया गया है उनमें कोई स्थिरता नहीं है। अब जब हम डाटा सेट को देखते है तो प्राथमिक नज़र ही हमें बताती है कि डाटा सेट में कुछ समस्या है।**

**जब भी हम डाटा को मापते हैं या एक वेरिएबल को मापते हैं जिसकी इकाइयाँ हैं तो हमें सभी अवलोकन (observation) में उपयोग की जा रही इकाइयों के बारे में सुसंगत (consistent) होना चाहिए। और इसी वजह से हम कहते हैं कि कॉलम वेरिएबल का प्रतिनिधित्व करते हैं और प्रत्येक वेरिएबल के लिए एक ही प्रकार का मान दर्ज किया जाना चाहिए। मान (value) के प्रकार से हमारा क्या मतलब है? इसको समझने के लिए उदाहरण के तौर पे हमारे डाटा सेट में यहां जन्म तिथि दी है जो 17 फरवरी 2003 है।**

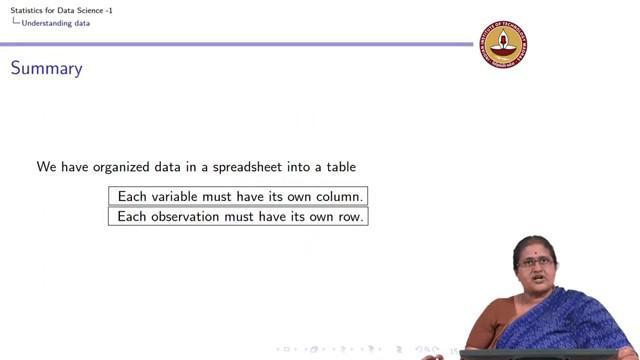
**(स्लाइड समय देखें: 29:01)**



**अब मान लीजिए कुछ लोगों के लिए हम ये जन्म तिथि बदलना चाहे जैसे हम इसे 19 साल या 21 साल 3 महीने करना चाहे तो तकनीकी रूप से कहें तो ये दोनों एक बात को ही कैप्चर कर रहे हैं, अगर हमे जन्म की तारीख पता है तो हम गणना कर पता कर सकते हैं कि उम्र क्या है; और अगर हमे उम्र पता है तो हम गणना कर पता कर सकते हैं की जन्म की तारीख क्या है। लेकिन यह डाटा का प्रतिनिधित्व करने का सही तरीका नहीं हैं।**

**इसलिए जब हम विशेष रूप से प्रारूप (format) में डाटा की गणना या संग्रह कर रहे हों तो हमे सुनिश्चित करने की आवश्यकता है कि पंक्तियां (rows) प्रत्येक केस का प्रतिनिधित्व कर रही हो और कॉलम, वेरिएबल का प्रतिनिधित्व कर रहे हो और प्रत्येक वेरिएबल के लिए हम समान प्रकार के मान को सुनिश्चित करे। एक और उदाहरण हम आपको यहाँ देते हैं जैसे हमारे पास यहाँ कक्षा में चार सौ चौरासी (484) अंक दिये हुए हो तो हम अचानक यह नहीं कह सकते कि यह बानवे (92) प्रतिशत या कुछ और है, भले ही तकनीकी रूप से यह भी एक मूल्यांकन का तरीका है।**

**(स्लाइड समय देखें: 30:28)**



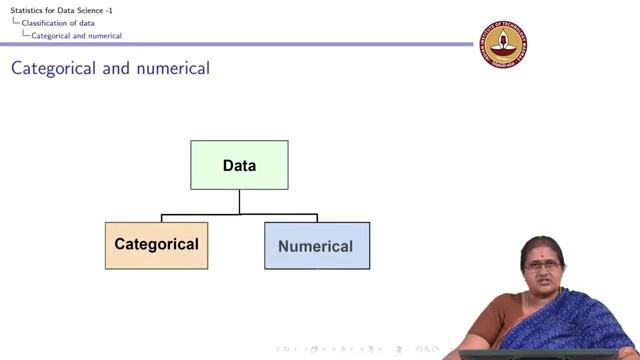
**तो यह एक चीज है जिसका हमें हर समय ध्यान रखने की जरूरत है; इसलिए, जब हम डाटा सेट के बारे में बात करते हैं तो हमे बहुत सावधान रहने की आवश्यकता है और हमे यह भी समझने की आवश्यकता है कि हमारे पास जो डाटा सेट है उसमे प्रत्येक वेरिएबल का अपना कॉलम है और हमने पहले ही परिभाषित कर दिया है कि वेरिएबल क्या है।**

**यदि चर यानि वेरिएबल की इकाइयाँ हैं, तो प्रत्येक अवलोकन (observation) की अपनी पंक्ति होगी और प्रत्येक वेरिएबल को प्रत्येक अवलोकन के लिए मापा जाएगा जहां इकाइयाँ सुसंगत (consistent) होंगी। उदाहरण के लिए अगर हमारे पास ऊंचाई की इकाई लेने वाला अवलोकन है तो यह नहीं हो सकता की एक अवलोकन हम सेंटीमीटर में ले और दूसरा अवलोकन फीट मे दिया जाए। हमे प्रत्येक अवलोकन के लिए एक ही प्रकार का मान लेने वाली परिवर्तनीय ऊंचाई की आवश्यकता होगी।**

**तो इस पाठ्यक्रम मे इस स्तर तक हमने जाना कि डाटा क्या है और डाटा सेट क्या है। डाटा सेट जिसे एक स्प्रेडशीट में व्यवस्थित किया जाएगा। हमने यहां कुछ उदाहरण भी प्रस्तुत किए। यह एक स्कूल** का **डाटा है जो स्प्रेडशीट में व्यवस्थित है, यह अस्पताल का डाटा है जिसे फिर से एक स्प्रेडशीट में व्यवस्थित किया गया है, ये खिलाड़ियों का डाटा हैं जो स्प्रेडशीट में व्यवस्थित हैं। हमें जो समझने और याद रखने की आवश्यकता है वह यह है कि कॉलम वेरिएबल का प्रतिनिधित्व करते हैं और रौज़ यानि पंक्तियाँ केस या अवलोकन का प्रतिनिधित्व करती हैं।**

**प्रत्येक अवलोकन के लिए हम चिह्नित करते हैं कि परिवर्तनीय मान क्या है। यदि वेरिएबल मान उपलब्ध नहीं है, तो हम इसे उपलब्ध नहीं है, के रूप मे नोट करते हैं; हम उस अवलोकन के लिए उस वेरिएबल की अनुपलब्धता को दर्शाते हैं। इस समय हमारे पास विश्लेषण के लिए एक डाटा सेट उपलब्ध है। हमारा अगला कदम होगा जहां हमें यह समझने की जरूरत होगी कि हम इस वेरिएबल यानि चर के बारे में क्या जानते हैं, हम इन अवलोकनों को कैसे वर्गीकृत करेंगे; इस बात पर हम आगे चर्चा करेंगे।**

**(स्लाइड समय देखें: 32:34)**



**GLOSSARY**

|  |  |  |
| --- | --- | --- |
| **ENGLISH WORD** | **WORD** | **MEANING** |
| **Statistical Analysis** | **स्टेटिस्टिक्ल एनालिसिस** | **सांख्यिकीय विश्लेषण** |
| **Process** | **प्रोसैस** | **संसाधित** |
| **Characteristics** | **कैरक्‌टरिस्‌टिक्‌** | **विशेषताएँ** |
| **Attribute** | **एट्रिब्यूट** | **गुण** |
| **Consistent** | **कंसिस्टेंट** | **सुसंगत** |
| **Observation** | **ऑब्जरवेशन** | **अवलोकन** |
| **Structured** | **स्ट्रक्चर्ड** | **संरचित** |