**डाटा साइंस के लिए स्टेटिस्टिक्स -1**

**प्रो. उषा मोहन**

**प्रबंधन अध्ययन विभाग**

**भारतीय प्रौद्योगिकी संस्थान, मद्रास**

**व्याख्यान संख्या 2.4**

**केटेगोरियल डेटा का वर्णन - डेटा रेखांकन(graphing) करने का सर्वोत्तम अभ्यास – 2**

(स्लाइड समय देखें: 00:14)



अब जब भी हम डेटा प्रदर्शित करने के लिए आते हैं, तो एक मूलभूत नियम होता है जिसे क्षेत्र सिद्धांत(area principle) कहा जाता है। अब, यह क्षेत्र सिद्धांत वास्तव में क्या है? क्षेत्र सिद्धांत बताता है कि ग्राफ़ के एक हिस्से द्वारा कब्जा कर लिया गया क्षेत्र उस डेटा की मात्रा के अनुरूप होना चाहिए जो इसे दर्शाता है। इससे हमारा क्या तात्पर्य है?

(स्लाइड समय देखें: 00:42)



पहली बात यह है कि जब मेरे पास चार्ट होते हैं जो ध्यान आकर्षित करने के लिए इस्तेमाल जाते हैं। उदाहरण के लिए, आइए अब इस विशेष चार्ट को देखें, यह चार्ट हमें संयुक्त राज्य में निर्यात की जाने वाली कुल शराब के बारे में बताता है जिसका मूल्य लाखों डॉलर में है। अब यह एक इन्फोग्राफिक(infographic) की तरह है, लेकिन यह क्या बताता है? यह बताता है कि यूनाइटेड किंगडम लगभग 150.3 है, कनाडा 146.8 है, जापान 82.8 है, इटली 42.5 है और नीदरलैंड 34.4 है।

आप तुरंत देख सकते तीहैं कि कोई आधार रेखा नहीं है, यह लेबल बॉक्स के ऊपर बोतलें दिखाती है। अब डिब्बे भी एक समान आकार के नहीं हैं, वे विभिन्न आकार के हैं। इसलिए, जब आप इस चार्ट को देखते हैं तो यह कुछ भी नहीं बताता है जो वास्तव में मैं इसका अर्थ निकाल सकती हूं। मुझे जो कुछ भी बताना है, वह इन देशों में से प्रत्येक को निर्यात की जाने वाली कुल अमेरिकी शराब है और मैं इसे एक बार चार्ट का उपयोग करके बना सकती हूं जहां मेरी श्रेणियां फिर से यूनाइटेड किंगडम, कनाडा, जापान, इटली और नीदरलैंड हैं।

मैंने अपनी प्रत्येक श्रेणी को लेबल किया है। मैं देख सकती हूं कि यूनाइटेड किंगडम का मूल्य 150.3 मिलियन डॉलर है, कनाडा लगभग 146.8 मिलियन डॉलर, जापान के लिए 82.8, इटली के लिए 42.5 और नीदरलैंड के लिए 34.4 है। अब दायीं ओर का चार्ट उसी का पालन करता है जिसे हम क्षेत्र सिद्धांत(area principle) कहते हैं। यह सटीक है, इसकी एक आधार रेखा(base line) है। तो, आप देख सकते हैं कि यह इस चार्ट की आधार (base line)है।

यह वास्तव में प्रत्येक देश के बार्स की चौड़ाई को बराबर बनाने के लिए के अनुरूप है और मेरे पास एक लंबवत स्केल(vertical scale) हो सकता है और लंबवत पैमाने पर मेरे पास वह मान है जो दिया गया है और इसलिए यह उस क्षेत्र सिद्धांत(area principle) का पालन करता है जहां ग्राफ द्वारा क्षेत्र पर अधिकृत कर लिया गया है |ग्राफ द्वारा अधिकृत कर लिया गया क्षेत्र क्या है? यह क्षेत्र है और यह क्षेत्र प्रस्तुत किए जा रहे डेटा के समानुपाती है।

(स्लाइड समय देखें: 03:21)



अगले तरीके से लोग ग्राफ़ के साथ गुमराह करते हैं, जिसे हम खण्डित ग्राफ़ के रूप में कहते हैं। अब एक खण्डित ग्राफ क्या है? अब मैं एक बात दिखाती हूं कि जहां एक बार चार्ट की आधार रेखा 0 पर नहीं है। बार चार्ट के आधार रेखा 0 पर नहीं है से हमारा क्या मतलब है? अब मैं आपको यह ग्राफ दिखाती हूँ ,फिर से यह एक क्षेत्रीय वार वितरण (regional wise distribution) है। मेरे चार क्षेत्र दक्षिण, पश्चिम, उत्तर और पूर्व हैं।

अब जब मैंने यह डेटा और यह ग्राफ आपके सामने पेश किया, तो आप तुरंत देख सकते हैं कि 500 लोग हैं और मेरे पास डेटा है कि 500 लोग किस क्षेत्र से आते हैं, वह डेटा है। अब जब मैं इस ग्राफ को चित्रित करती हूं या दिखाती हूं, तो तत्काल प्रतिक्रिया या कोई व्यक्ति जो इस डेटा को बिना पहुंच देखे देखता है तो यह कल्पना करता है कि दक्षिण और उत्तर के लोग पश्चिम या पूर्व के लोगों की तुलना में बहुत अधिक हैं।

यह ग्राफ़ यही संदेश देता है, लेकिन जब हम डेटा को अधिक ध्यान से देखते हैं तो आप देखते हैं कि यह आधार रेखा है और यह 0 से शुरू होता है जबकि यहां यह 75 से शुरू होता है। अब ये दोनों ग्राफ वास्तव में एक ही डेटा हैं। समान डेटा से मेरा क्या मतलब है? उनमें से प्रत्येक 500 छात्र बताते हैं कि वे किस राज्य, क्षेत्र से आते हैं चाहे वह दक्षिण हो, चाहे वह पश्चिम हो, चाहे वह उत्तर हो या पूर्व।

जब आप इस ग्राफ को देखते हैं, तो आपको लगता है कि दक्षिण और उत्तर वे क्षेत्र हैं जहां से अधिकांश छात्र आते हैं। भले ही बहुमत हो। ऐसा लगता है कि उत्तर और पूर्व, उत्तर और दक्षिण से वितरण के लिए नगण्य(negligible) हैं, जबकि यह ग्राफ एक अलग कहानी बताता है। तो, बायां ग्राफ या यह ग्राफ संख्या को बढ़ा-चढ़ाकर पेश करता है जबकि यह ग्राफ 0 पर समान आधार रेखा दिखाता है।.

अब, यही अनुशंसा (recommendation) की जाती है क्योंकि यह वह जगह है जहां डेटा अखंडता (data integrity) को बनाए रखा जाता है और वास्तविक कहानी कहा जाता है। अब यह कहाँ मायने रखता है, अब जब आप विकास और सब कुछ दिखा रहे हैं और आप ग्राफ को काट-छांट कर रहे हैं या आप वर्तमान डेटा में बदलाव करते हैं या आप गुमराह करते हैं, तो लोग वास्तव में इसके लिए एक गलत कहानी का श्रेय देंगे, भले ही दोनों एक ही डेटा का प्रतिनिधित्व करते हों। आप देखते हैं कि यह एक अलग कहानी बताता है जो दाईं ओर वाला कहता है।

तो, दूसरी बात यह है कि जब भी आप ग्राफ़ को खण्डित करते हैं तो जानकारी का नुकसान होता है और यह उल्लंघन है। इससे बचना चाहिए।

(स्लाइड समय देखें: 06:42)



अब कुछ पाठ्यपुस्तकें और कुछ लोग कहते हैं कि जब भी आपके पास एक खंडित ग्राफ या कुछ और हो, तो आपको Yअक्ष(Y axis) विराम का परिचय देना चाहिए। इसलिए, यदि आप शुरू कर रहे हैं और आप धुरी तोड़ रहे हैं तो आप इसे या तो ऐसा करके पेश कर सकते हैं जहां मैं कह रही हूं कि मैं अपने Yअक्ष(Y axis) में एक ब्रेक पेश कर रही हूं या आप यह दिखा सकते हैं कि यह मैं एक ब्रेक 0 से शुरू नहीं कर रही हूं और मैं एक उच्च मूल्य से शुरू कर रही हूँ।

उदाहरण के लिए, अगर मैं 75 से शुरू करना चाहती हूं तो मुझे संकेत देना चाहिए कि मैं ग्राफ को 75 पर स्थानांतरित करने के बजाय 75 से शुरू कर रही हूं। तो, यह गलत है जबकि इसके बारे में जाने का यह सही तरीका है। इसलिए, जब भी आप Y अक्ष(Y axis) में परिवर्तन या हेरफेर कर रहे हों, तो पाठक को संकेत दें कि आपने हेरफेर किया है या आपने Y अक्ष(Y axis) विराम(break) की शुरुआत की है।

**Glossary**

|  |  |  |
| --- | --- | --- |
| **Words** | **Transliteration** | **Translation/Meaning** |
| Area principle | एरिया प्रिंसिपल | क्षेत्र सिद्धांत- ग्राफ़ का क्षेत्रफल उस डेटा के परिमाण के बराबर होना चाहिए जिसे वह प्रदर्शित कर रहा है। |
| Axis | एक्सिस | अक्ष- निर्देशांक की माप के लिए एक निश्चित संदर्भ रेखा। |
| Base line | बेस लाइन | आधार रेखा- तुलना के लिए उपयोग किया जाने वाला न्यूनतम या प्रारंभिक बिंदु। |
| Data integrity | डेटा इंटीग्रिटी | डेटा अखंडता- अपने पूरे जीवनचक्र में डेटा की सटीकता, पूर्णता और विश्वसनीयता | |
| Info graphic | इन्फो ग्राफिक | आलेख जानकारी- सूचना या डेटा का एक दृश्य प्रतिनिधित्व, उदाहरण के लिए चार्ट या आरेख के रूप में। |
| Vertical scale | वर्टिकल स्केल | लंबवत स्तर |