Introduction : Measures of Dispersion

ဒီသင်ခန်းစာမှာ Measures of Dispersion အကြောင်းကို ပြောသွားပါမယ်။ ဒီ Measures of Dispersion ကလည်း ပြီးခဲ့တဲ့သင်ခန်းစာမှာတုန်းက ပြောပြခဲ့တဲ့ Measures of Central Tendency လိုပဲ ကိုယ့်ရဲ့ဒေတာတွေကို ပိုမိုနားလည်အောင်ပြုလုပ်ပေးနိုင်တဲ့ အရေးကြီးတဲ့အကြောင်းအရာတစ်ခု ဖြစ်ပါတယ်။

Dispersion ဆိုတာက ကိုယ့်ဒေတာရဲ့အလယ်တည့်တည့်ကို ကြည့်တာမဟုတ်ဘဲ ကိုယ့်ရဲ့ ဒေတာထဲမှာပါတဲ့ တန်ဖိုးတွေက တစ်ခုနဲ့တစ်ခု ဘယ်လောက်ကွာဝေးပြီး ပြန့်နှံ့တည်ရှိနေသလဲဆိုတာကို ကြည့်တာဖြစ်ပါတယ်။

ဒီလိုကြည့်တဲ့အခါမှာ Customer တွေရဲ့ဝယ်ယူမှုတွေ၊ ကိုယ့်ဝက်ဘ်ဆိုဒ်မှာ Customer တွေ ဘယ်လောက်ကြာကြာနေသလဲ ဘယ်လိုတွေပြုမူကြသလဲဆိုတဲ့အချက်အလက်တွေရဲ့ **ကွဲလွဲမှု(variance)** ကို နားလည်စေမှာ ဖြစ်ပါတယ်။

အခု Measures of Dispersion သင်ခန်းစာမှာဆိုရင် Variance အကြောင်းကို အရင်ဆုံးပြောပြသွားမှာပါ။ Variance ကိုအသုံးပြုခြင်းအားဖြင့် ကိုယ့်ရဲ့ အန္တရာယ် (risk) ကိုဘယ်လိုပိုနားလည်စေမလဲဆိုတာရယ်၊ ကိုယ့်ရဲ့ ဘယ် customer (audience) ကို ပိုပြီးဦးစားပေးရမလဲဆိုတာတွေရယ်ကို ရှင်းပြပေးသွားမှာပါ။

နောက်တစ်ခုအနေနဲ့ကတော့ Standard Deviation (စံကွဲလွဲမှု) ပါ။ Standard Deviation ကိုအသုံးပြုပြီးတော့ Variants (ကွဲလွဲမှုတွေ)ကို ဘယ်လိုတိုင်းတာနိုင်မလဲ? ကိုယ့်ရဲ့ Customer တွေနဲ့ လုပ်ငန်းအကြောင်းကို ပိုပြီးနားလည်နိုင်အောင် Standard Deviation က ဘယ်လိုကူညီနိုင်မလဲဆိုတာတွေကို ရှင်းပြသွားမှာ ဖြစ်ပါတယ်။

နောက်ဆုံးအနေနဲ့ကတော့ Z-score အကြောင်းကို ပြောသွားပါမယ်။ ကိုယ့်ရဲ့ Data Set ထဲက အချို့သောတန်ဖိုးတွေကို အကဲဖြတ်ဖို့အတွက် Z-score ကိုဘယ်လိုအသုံးပြုရမယ်ဆိုတာရယ်၊ ဒီလိုအသုံးပြုခြင်းအားဖြင့် ကိုယ့်ရဲ့ Sale & Marketing Project တွေမှာ ဘယ်လိုပိုပြီးပစ်မှတ်ထားနိုင်မလဲဆိုတာကို ရှင်းပြသွားမှာပါ။

အားလုံးပြီးသွားတဲ့နောက်မှာတော့ Excel / Spreadsheet ထဲမှာ standard deviation ကိုဘယ်လိုတွက်ရမလဲဆိုတာ ရှင်းပြပေးသွားပါမယ်။ ကဲ ဒီတော့ ဆက်လိုက်ကြရအောင်ပါ။

Variance in Data Analytics

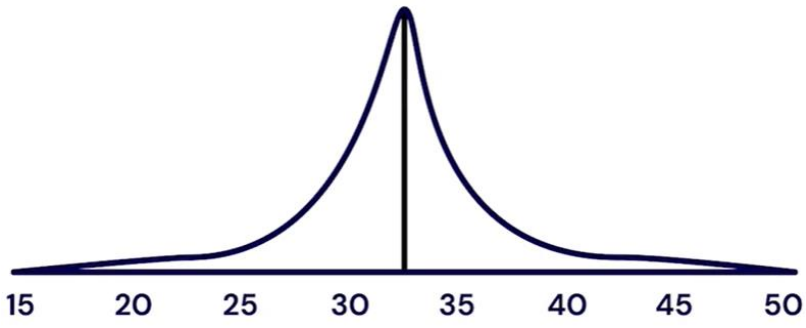
Variance ဆိုတာက Data Analytics မှာအရေးကြီးတဲ့အကြောင်းအရာတစ်ခုဖြစ်ပါတယ်။ ကိုယ့်ရဲ့ Customer တွေကို ပိုပြီးနားလည်လာစေဖို့ Variance ကို အသုံးပြုနိုင်ပါတယ်။

ဒီတော့ Varaiance ဆိုတာ ဘာပါလဲ? **Variance ဆိုတာက ကိုယ့် Data Set ထဲမှာရှိတဲ့ ဒေတာပွိုင့်တွေက Data Set ရဲ့ Mean/Average (ပျမ်းမျှတန်ဖိုး) ကနေ ဘယ်လောက်အကွာအဝေးတွေမှာ ပြန့်နှံ့တည်ရှိနေသလဲ ဆိုတာကို ဖော်ပြပေးတဲ့ ကိန်းဂဏန်းတစ်ခုပဲ ဖြစ်ပါတယ်။**

Data Analytics မှာ Variance ကို နည်းလမ်းအမျိုးမျိုးနဲ့ အသုံးချကြပါတယ်။ ဥပမာ - အန္တရာယ်ဖြစ်နိုင်ချေ ဆန်းစစ်မှု(Risk Analysis) လိုမျိုးမှာ Variance က အရမ်းကို အရေးပါပါတယ်။ Variance တန်ဖိုးက အရမ်းကို မြင့်မားနေမယ်ဆိုရင် ထွက်ပေါ်လာမယ့်ရလဒ်(outcome)ကို မှန်မှန်ကန်ကန် ကြိုတင်ခန့်မှန်းနိုင်မှာ မဟုတ်ပါဘူး။ အဲ့ဒီအတွက်ကြောင့် ပိုပြီးမြင့်မားတဲ့ Risk ကို ဦးတည်သွားစေပါတယ်။

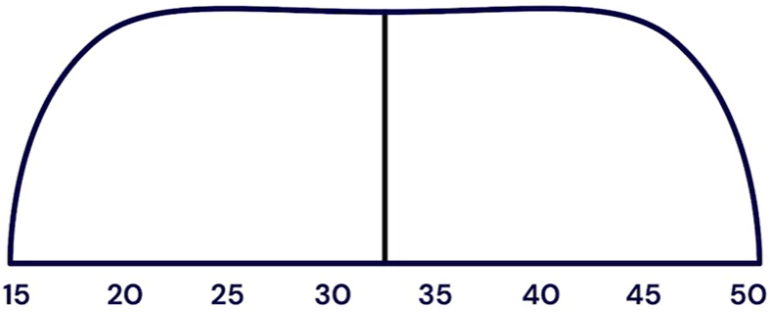
Variance ကို ဒေတာရဲ့စိတ်ချယုံကြည်ရမှု(data reliability) ကို စိစစ်ဖို့အတွက်လည်း အသုံးပြုနိုင်ပါတယ်။ တနည်းအားဖြင့်ပြောမယ်ဆိုရင် ကိုယ့်ဒေတာကပြောနေတာက တကယ်ယုံလို့ရော ရရဲ့လားဆိုတာကို စစ်ဆေးကြည့်တာဖြစ်ပါတယ်။

နောက်ထပ် Variance ကိုအသုံးပြုနိုင်တာကတော့ ဥပမာ Marketing Field မှာဆိုရင် ကိုယ့်ရဲ့ Targeted Marketing Campaign က ကိုယ်ပစ်မှတ်ထားတဲ့ Customer/Audience တွေဆီကို အောင်အောင်မြင်မြင်ရောက်ရှိရဲ့လားဆိုတာကို အကဲဖြတ်ဖို့အတွက် အသုံးပြုနိုင်ပါတယ်။ ပိုမြင်သာအောင် ဥပမာတစ်ခု ပြပါမယ်။ ကျွန်တော်တို့က အသက် ၃၀ ကနေ ၃၅ နှစ်အတွင်းမှာရှိတဲ့ Customer တွေကိုပစ်မှတ်ထားပြီး Marketing Campaign တစ်ခုလုပ်တယ်ဆိုပါစို့။ Campaign ပြီးသွားတဲ့အချိန်မှာ ကိုယ့်ရဲ့ Campaign က အောင်မြင်တယ်၊ မအောင်မြင်ဘူးဆိုတာကို သိရှိဖို့ ကိုယ့်ဆီကနေပစ္စည်းဝယ်သွားတဲ့ Customer တွေရဲ့အသက်ကို လေ့လာကြည့်ပါမယ်။ Campaign က အောင်မြင်တယ်ဆိုရင် ပျမ်းမျှအသက်တန်ဖိုးက ၃၀ နဲ့ ၃၅ ကြားမှာ ရှိနေသင့်ပါတယ်။



ပစ္စည်းဝယ်ယူသွားတဲ့ Customer တွေရဲ့အသက်က ၃၀ ထက် အနည်းငယ်ပိုငယ်နေတာ၊ ၃၅ ထက် အနည်းငယ် ပိုကြီးနေတာမျိုးတွေတော့ ပါလာနိုင်ပါတယ်။ ဒါပေမယ့်လည်း ကျွန်တော်တို့ရဲ့ variance - ကွဲလွဲမှု (The amount of spread) က နှိုင်းရအားဖြင့်တော့ နည်းနေသင့်ပါတယ်။ တနည်းအားဖြင့် အသက်တန်ဖိုးတွေဟာ

Mean ကနေ အများကြီးကွာမနေသင့်ပါဘူး။ အကယ်၍ variance က မြင့်မားနေခဲ့မယ်ဆိုရင် ပစ္စည်းဝယ်ယူသွားတဲ့ Customers တွေရဲ့အသက်ဟာ အရမ်းကိုကွဲကွဲပြားပြား ပြန့်နှံ့တည်ရှိနေတာဖြစ်ပြီး ကိုယ်ပစ်မှတ်ထားတဲ့ Customer/Audience ဆီကို မရောက်ဘူးလို့ ပြောနိုင်ပါတယ်။



ဒါပေမယ့် အခုလိုမရောက်ဘူးဆိုတာကို သိလိုက်ရတဲ့အချက်(Insight) ကိုအသုံးပြုပြီးတော့ နောက်တစ်ကြိမ် Marketing Campaign လုပ်တဲ့အခါမှာ ကိုယ်ပစ်မှတ်ထားတဲ့အသက် ၃၀-၃၅ ရှိတဲ့ Customer တွေဆီကို ပိုပြီးတော့ရောက်ရှိစေဖို့ Marketing လုပ်တဲ့နည်းလမ်းတွေကို ပြန်လည်ပြင်ဆင်နိုင်စေမှာ ဖြစ်ပါတယ်။ ဒါဟာ Data Analytics မှာ variance ကိုအသုံးပြုတဲ့နည်းလမ်းတွေထဲက တစ်ခု ဖြစ်ပါတယ်။

အခုဆိုရင် ကျွန်တော်တို့က variance ရဲ့အရေးပါပုံကို နားလည်သွားပါပြီ။ ဒီတော့ မေးစရာရှိတာက variance ကို ဘယ်လိုရှာမလဲဆိုတဲ့မေးခွန်းဖြစ်ပါတယ်။ ကျွန်တော်တို့ရဲ့ဒေတာက ဘယ်လိုပြန့်နှံ့တည်ရှိနေသလဲဆိုတာကို သိရှိအောင်လုပ်ဖို့ နည်းလမ်းတွေအများကြီး ရှိပါတယ်။

ပထမဆုံးနဲ့ အရိုးရှင်းဆုံးနည်းလမ်းက **range** ဖြစ်ပါတယ်။ range ဆိုတာက ကိုယ့်ဒေတာရဲ့အမြင့်ဆုံးတန်ဖိုးနဲ့ အနိမ့်ဆုံးတန်ဖိုးကြားမှာ ဘယ်လောက်ကွာဟသလဲဆိုတာကို ပြသပေးတာ ဖြစ်ပါတယ်။ range ကို ရှာဖို့က အမြင့်ဆုံးတန်ဖိုးကနေ အနိမ့်ဆုံးတန်ဖိုးကို နှုတ်ပေးလိုက်ရုံပါပဲ။ ဥပမာ ကျွန်တော်တို့ရဲ့ ဒေတာထဲမှာ 1 ကနေ 100 အထိ တန်ဖိုးတွေပါနေတယ်ဆိုပါဆို့။ ဒါဆိုရင် 100-1=99 ဖြစ်တဲ့အတွက် range တန်ဖိုးဟာ 99 ဖြစ်ပါတယ်။

Neko ကုမ္ပဏီနဲ့ ဥပမာတစ်ခုကို ကြည့်လိုက်ရအောင်။ ကုမ္ပဏီမှာ ပြီးခဲ့တဲ့လက Customers တွေရဲ့ ဝယ်ယူမှုပမာဏတွေကို စာရင်းလုပ်ထားပါတယ်။

Date	Purchase Amount	
7/1	$50.00	Mean
7/2	$53.00	$50.00
7/3	$47.00	
7/4	$49.00	
7/5	$51.00	
7/6	$50.00	
7/7	$49.00	
7/8	$48.00	
7/9	$51.00	
7/10	$53.00	
7/11	$47.00	
7/12	$48.00	
7/13	$50.00	
7/14	$49.00	
7/15	$48.00	
7/16	$47.00	
7/17	$50.00	
7/18	$50.00	
7/19	$47.00	
7/20	$52.00	
7/21	$51.00	
7/22	$50.00	
7/23	$49.00	
7/24	$53.00	
7/25	$51.00	
7/26	$50.00	
7/27	$52.00	
7/28	$50.00	
7/29	$50.00	
7/30	$53.00	
7/31	$52.00	

စာရင်းအရဆိုရင် ပျမ်းမျှဝယ်ယူမှုပမာဏက $50 ဖြစ်ပါတယ်။ ဒါပေမယ့် ကုမ္ပဏီက Customers တွေအားလုံးက $50 ဝန်းကျင်ပဲ ဝယ်ယူကြတာလား ဒါမှမဟုတ် အဲ့ဒီပမာဏထက် အများကြီးပိုပြီးဝယ်တာ ဒါမှမဟုတ် အများကြီးလျော့ပြီးဝယ်တာ ရှိသလားဆိုတာ သိချင်နေပါတယ်။

ဒီမေးခွန်းကိုဖြေဖို့အတွက် range တန်ဖိုးကို တစ်ချက်ကြည့်လိုက်ရုံနဲ့ သိနိုင်ပါတယ်။ Dataset ထဲမှာရှိတဲ့ အကြီးဆုံးတန်ဖိုးက $53 ဖြစ်ပြီး အသေးဆုံးတန်ဖိုးက $47 ဖြစ်မယ်ဆိုရင် 53 – 47 = 6 ဖြစ်တဲ့အတွက် range တန်ဖိုးဟာ $6 ဖြစ်ပါတယ်။ variance တန်ဖိုးက နည်းတာကို တွေ့ရပါတယ်။ ဒါကြောင့် ကုမ္ပဏီရဲ့ Customer တွေဟာ $50 ဝန်းကျင်ပဲ ဝယ်ယူကြတယ်လို့ ကောက်ချက်ချနိုင်ပါတယ်။

ဒါပေမယ့် Customer တွေရဲ့အနည်းဆုံးဝယ်ယူမှုက $1 ဖြစ်ပြီး အများဆုံးဝယ်ယူမှုက $99 ဖြစ်တယ်ဆိုရင်ရော?

Date	Purchase Amount	
7/1	$38.00	Mean
7/2	$53.00	$50.00
7/3	$1.00	
7/4	$15.00	
7/5	$50.00	
7/6	$68.00	
7/7	$75.00	
7/8	$48.00	
7/9	$51.00	
7/10	$99.00	
7/11	$47.00	
7/12	$48.00	
7/13	$50.00	
7/14	$60.00	
7/15	$74.00	
7/16	$18.00	
7/17	$25.00	
7/18	$73.00	
7/19	$47.00	
7/20	$52.00	
7/21	$51.00	
7/22	$50.00	
7/23	$23.00	
7/24	$93.00	
7/25	$51.00	
7/26	$26.00	
7/27	$52.00	
7/28	$43.00	
7/29	$50.00	
7/30	$87.00	
7/31	$32.00	

99 – 1 = 98 ဖြစ်တဲ့အတွက် range တန်ဖိုးဟာ $98 ဖြစ်ပါလိမ့်မယ်။ အောက်ပါပုံကိုကြည့်မယ်ဆိုရင် variance တန်ဖိုးက မြင့်နေတာကို တွေ့ရမှာပါ။

Dataset အရ Mean တန်ဖိုးဟာ $50 ဖြစ်နေဆဲဆိုပေမယ့် variance က အများကြီး မြင့်နေပါတယ်။ ဒါကြောင့် Customer တွေဟာ ပျမ်းမျှတန်ဖိုးဝန်းကျင်မှာတင် ဝယ်ယူတာမျိုးမဟုတ်ဘဲ ပမာဏအမျိုးမျိုးနဲ့ ဝယ်ယူကြတယ်လို့ ကောက်ချက်ချရမှာ ဖြစ်ပါတယ်။

Variance လိုမျိုး Descriptive Statistics ဟာ ကိုယ့်ဒေတာရဲ့အခြေခံအချက်အလက်တွေကို မြန်မြန်ဆန်ဆန် လွယ်လွယ်ကူကူ နားလည်လာအောင် ပြုလုပ်ပေးနိုင်ပါတယ်။ အခုပြခဲ့တဲ့နမူနာအရဆိုရင် ကျွန်တော်တို့ဟာ Customers တွေရဲ့ ဝယ်ယူမှုပုံစံ (Purchase Behavior) ကို ပိုပြီးနားလည်သွားပါပြီ။

Data Analyst တစ်ယောက်အနေနဲ့ ကိုယ့်ရဲ့ဒေတာကနေထုတ်ပေးနိုင်တဲ့ Insights အမျိုးမျိုးကို ပိုပြီးနားလည်ထားလေလေ ပိုကောင်းတဲ့မေးခွန်းတွေကို ပိုပြီးထုတ်နိုင်လာလေလေပါပဲ။ Variance ဟာ Data Analysis process မှာ အရေးကြီးတဲ့အပိုင်းကနေ ပါဝင်နေပြီးတော့ ကျွန်တော်တို့ကို ပိုကောင်းတဲ့ Data Analyst တွေ ဖြစ်လာဖို့ ကူညီပေးနိုင်ပါတယ်။