ফলিত পরিসংখ্যান ও ডেটা সায়েন্স (লেখা চলছে)

এনায়েতুর রহীম

2019-08-14

সূচীপত্র

অধ্যায় ১

পাঠকের প্রতি

এই বইটিতে এখনও লেখার কাজ চলছে। এবং যখনই নতুন লেখা যোগ করা হবে তখনই এই সাইটটি আপডেট করা হবে। তার মানে হল চ্যাপ্টারগুলোর ক্রম যেকোন সময় পরিবর্তিত হতে পারে। সেই সাথে সেকশন ও সাবসেকশনের নম্বরও বদলে যেতে পারে। তাই পাঠকের কাছে অনুরোধ থাকবে যেন কোন অধ্যায়কে বুকমার্ক না করতে। বরং পুরো বইটিকে বুকমার্ক করতে।

১.১ ভার্সন সমূহ (ই-বুক, অনলাইন)

এই বইটি ফ্রিতে অনলাইনে পাওয়া যাবে। সাথে ই-বুক ফরম্যাট দেয়ার ইচ্ছে রয়েছে। ই-বইয়ের পরীক্ষামূলক একটি ভার্সন অচিরেই প্রকাশ করব। অনলাইন ভার্সনের ওয়েবসাইটের ঠিকানা https://asds.dataskool.org। সাইটে গিয়ে ই-বুক ফরম্যাট ডাউনলোড করা যাবে (যখন রেডি হবে)। ই-বুক পড়ার জন্য যেকোন এ্যাপ দিলেই হবে না। কারণ এই বইয়ে গাণিতিক ফরমুলা ও চিহ্ন আছে যা পড়ার জন্য GitdenReader নামক এ্যাপ দরকার হবে। এ্যান্ড্রয়েড ও এপল এ্যাপ স্টোর থেকে এটি ফ্রি পাওয়া যাবে। অন্য ই-রিডার ব্যবহার করা যাবে তবে গাণিতিক ফরমুলা থাকায় সব রিডার দিয়ে ফরমুলা ও গাণিতিক চিহ্ন ঠিকমত পড়া যাবে না।

১.২ প্রশ্ন ও পরামর্শ

বইটিতে কোন ভুল ভ্রান্তি পেলে তা আমি জরুরীভিত্তিতে শোধরানোর চেষ্টা থাকবে। কোন ভুল পেলে কিংবা কোন পরামর্শ থাকলে বা কোন টপিকের অনুরোধ নীচের ঠিকানায় জানানোর জন্য অনুরোধ করছি।

https://github.com/raheems/asds/issues

এর জন্য গিটহাবের একাউন্ট একাউন্ট লাগবে। একাউন্ট করা খুবই সহজ। ওখানে গেলেই একাউন্ট করার অপশন দিবে। একাউন্ট করার পর লগিন করে আপনার পরামর্শ জমা দিন; আমি যথা সম্ভব দ্রুত উত্তর দেয়ার চেষ্টা করব।

১.৩ পরিকল্পনা

বইটিতে যেসব অধ্যায় থাকবে বলে ঠিক করেছি তার একটি তালিকা দিচ্ছি। আমি মোটামুটি নিশ্চিত যে এই তালিকা পরিবর্তিত হবে।

- ভূমিকা
- ডেটা কী ও দেখতে কেমন?
- ডেটা ভিজুয়ালাইজেশন
    - ডেটার শেইপ নিয়ে আলোচনা
- লোকেশন ও ভ্যারিয়েশনের পরিমাপ
    - মিন, মিডিয়ান, মোড
    - রেইন্জ, ভ্যারিয়্যান্স, স্ট্যান্ডার্ড ডিভিয়েশন
- ভ্যারিয়েবলের মধ্যে সম্পর্ক
    - বাইভ্যারিয়েট সম্পর্ক
    - মাল্টিভ্যারিয়েট সম্পর্ক
    - রিগ্রেশনের ধারণা
- প্রবাবিলিটি সম্পর্কে ধারণা
    - সেট থিউরি
    - ইভেন্ট
- রিগ্রেশন এনালিসিস
- এনালিসিস অব ভ্যারিয়েন্স

অধ্যায় ২

ডেটা ও ভ্যারিয়েবল

ডেটা সায়েন্স নিয়ে আলোচনার পর অনেকের হয়তো প্রশ্ন থাকবে যে ডেটা কী? ডেটা কোথায় থেকে আসে এবং সেগুলো কীভাবে কম্পিউটারে স্টোর করা হয়। কীভাবেই বা সেগুলো ব্যবহার করা হয়। ডেটা থেকে সিদ্ধান্ত গ্রহণ করে কীভাবে?

চলুন ডেটা এনালাইজ করার আগে আমরা ডেটা কী সে সম্পর্কে জানি।

> ডেটা হলো কোন ব্যক্তি বা বিষয়ের বৈশিষ্ট্য বা ক্যারেক্টারিসটিক যা সংখ্যা, চিহ্ন, ছবি, অডিও, ভিডিও, কিংবা লিখিত বা টেক্সট-আকারে কোন মাধ্যমে (যেমন মুদ্রণ বা কম্পিউটারে) সংরক্ষণ করা হয়। সংরক্ষিত ডেটা বা বৈশিষ্ট্যসমূহকে পরবর্তীতে বিশ্লেষণের মাধ্যেমে বিশেষ কোন কাজে লাগানোর উপযোগী করে যা পাওয়া যায় তাকে তথ্য (Information) বলে।

২.১ ডেটার উদাহরণ

২.১.১ ফেইসবুক লাইক সংখ্যা

আপনার ফেইসবুক পোস্টে লাইকের সংখ্যার কথা ধরা যাক। ধরা যাক আপনি গত এক মাসে মোট ৫০ টি পোস্ট করেছেন। প্রতিটি পোস্টে লাইক-এর সংখ্যা যদি আপনি গণনা করেন তাহলে আপনার পোস্টগুলো আপনার বন্ধুদের মাঝে কতটা আগ্রহ সৃষ্টি করেছে সেই বৈশিষ্ট্য সম্পর্কে একটা ধারণা আপনি পেতে পারেন। এই লাইকের সংখ্যাই ডেটা।

ধরি, লাইক সংখ্যাগুলো এরকম: 67, 41, 77, 30, 43, 59, 7, 37, 43, 42, 50, 48, 41, 70, 27, 43, 65, 50, 49, 56, 54, 48, 61, 56, 46, 28, 41, 33, 13, 41। এখানে ৩০ টি সংখ্যা আছে যেগুলো আপনার ৩০টি পোস্টের প্রতিটিতে কতগুলো লাইক দিয়েছে তা গণনা করেছেন। এই ডেটাকে নানা ভাবে সংরক্ষণ করা যেতে পারে। ফেইসবুক তাদের সার্ভারে অন্য সবার ডেটার সাথে আপনার এই ডেটাটিও সংরক্ষণ করে। ধরা যাক আপনি

সারনী ২.১: ফেইসবুক পোস্টে লাইক-এর সংখ্যা

likes
67
41
77
30
43
59
7
37
43
42

এই ডেটাগুলোকে আপনার কোন একটি কাজের জন্য সরক্ষণ করবেন। তার জন্য আমরা ডেটাকে সাধারণত ট্যাবুলার ফর্মে বা সারণি আকারে সাজাই যেটি দেখতে টেবিল ২.১-এর মতো হবে।

২.১.২ দৈনিক ইন্টারনেট ব্যবহারের পরিমাণ

আমরা প্রায় সবাই মোবাইল ইন্টারনেট প্যাকেজ ব্যবহার করি। আপনার দৈনিক কতটা ইন্টারনেট ব্যবহার করেন সেটি আপনার একটি বৈশিষ্ট্য যা ইন্টারনেট ব্যবহারে সাথে সম্পর্কযুক্ত। এই বৈশিষ্ট্যকে আমরা নানা ভাবে পরিমাপ করতে পারি। যার একটি হতে পারে আপনি কী পরিমাণ ইন্টারনেট ডেটা ব্যবহার করেন। ধরা যাক জানুয়ারি ২০১৮ থেকে অগাস্ট ২০১৮ পর্যন্ত প্রতি দিনের ইন্টারনেট ডেটা ব্যবহারের হিসাব রেকর্ড করা আছে।

ডেটার প্রথম দশ দিনের হিসাব টেবিল ২.২-এ দেয়া হল।

২.১.৩ মিরপুর টু মতিঝিল কমিউট টাইম

সকাল বেলা মিরপুর থেকে মতিঝিল যেতে যত সময় লাগে 2017-01-01 তারিখ থেকে 2018-08-31 তারিখ পর্যন্ত প্রতি দিনের হিসাব রেকর্ড করা হয়েছে। এটি একটি সিমুলেটেড ডেটাসেট এবং বাস্তবের সাথে এর মিল না থাকাই স্বাভাবিক। এখানে তিনটি মাধ্যম বিবেচনা করা হয়েছে - - বাস, উবার, এবং পাঠাও মটর সাইকেল সার্ভিস। ডেটার প্রথম ১০টি সারি টেবিল ২.৩- তে দেয়া হল।

সারনী ২.২: দৈনিক ইন্টারনেট ব্যবহারের পরিমাণ (মেগাবাইট)

dates	days	sex	usage
2018-01-01	Monday	Female	210.36
2018-01-02	Tuesday	Female	124.49
2018-01-03	Wednesday	Male	253.57
2018-01-04	Thursday	Female	99.51
2018-01-05	Friday	Male	129.80
2018-01-06	Saturday	Female	176.07
2018-01-07	Sunday	Male	62.37
2018-01-08	Monday	Female	114.35
2018-01-09	Tuesday	Female	129.22
2018-01-10	Wednesday	Female	125.63

২.২ ভ্যারিয়েবল ও প্রকারভেদ

ফেইসবুক লাইক ডেটাতে একটি মাত্র কলাম যার নাম দেখাচ্ছে likes। আর likes এর মান গুলো সব একই নয়, আলাদা আলাদা। likes- - কে বলে ভ্যারিয়েবল (variable) বা বাংলায় বলে চলক। পরিসংখ্যানে ভ্যারিয়েবল শব্দটি প্রায়ই শোনা যাবে। আমরা ভ্যারিয়বলের টেকনিক্যাল সংজ্ঞা দিতে চাই না। কারণ, বাস্তব ক্ষেত্রে সেটি জানা জরুরী নয়। সহজভাবে বলা যায়

> ভ্যারিয়েবল হলো ডেটা ফাইলের মধ্যে যে কলামগুলো থাকে সেগুলো। এদেরকে ভিন্ন ভিন্ন নাম দেয়া হয় কারণ ডেটার ভিতর দুটি ভ্যারিয়বলের নাম একই হতে পারে না।

আমরা এই ভ্যারিয়েবলের নাম অন্য কিছুও দিতে পারতাম। সাধারণত এমন নামই দেয়া হয় যা থেকে ভ্যারিয়েবল ও তার মান সম্পর্কে একটা ধারণা পাওয়া যায়। এই ডেটায় একটি মাত্র ভ্যারিয়বল আছে বলে এরকম ডেটাকে ইউনিভ্যারিয়েট ডেটা (univariate data) বলে। যে ডেটায় দুটি ভ্যারিয়েবল থাকবে তাকে বলবে বাইভ্যারিয়েট (bivariate) ডেটা, এবং দুয়ের অধিক ভ্যারিয়বল থাকলে সেই ডেটাকে মাল্টিভ্যারিয়েট (multivariate) ডেটা বলে।

ফেইসবুক লাইক ডেটাতে তিনটি ভ্যারিয়েবল আছে। সেগুলো হচ্ছে dates, days, sex, usage। যেহেতু দুয়ের অধিক ভ্যারিয়বল আছে সেহেতু এই ডেটাকে মাল্টিভ্যারিয়েট ডেটা বলা হবে।

আবার মিরপুর মতিঝিল কমিউট টাইম ডেটাতে ছয়টি ভ্যারিয়েবল আছে।

ভ্যারিয়েবল হতে পারে শুধু সংখ্যা, কিংবা তারিখ, কিংবা দিনের নাম (টেবিল ২.২ দ্রষ্টব্য)। কোন একটি ভ্যারিয়েবল কী রকম মান গ্রহণ করে তার উপর নির্ভর করে ভ্যারিয়েবলের ধরন। মূলত ভ্যারিয়বল দুই ধর-

সারনী ২.৩: মিরপুর থেকে মতিঝিল সকালে কমিউট সময় (ঘন্টায়)

dates	day_name	bus	uber	pathao	time_of_day
2017-01-01	Sunday	1.3	1.7	1.2	Morning
2017-01-02	Monday	2.9	1.8	1.9	Morning
2017-01-03	Tuesday	2.2	1.8	1.0	Morning
2017-01-04	Wednesday	2.0	1.2	1.1	Morning
2017-01-05	Thursday	3.5	3.0	1.5	Morning
2017-01-06	Friday	1.6	2.3	0.8	Morning
2017-01-07	Saturday	2.6	1.4	1.0	Morning
2017-01-08	Sunday	1.8	1.5	2.5	Morning
2017-01-09	Monday	2.2	2.6	1.5	Morning
2017-01-10	Tuesday	2.3	2.6	1.2	Morning

নের: কোয়ান্টিটেটিভ (quantitative) বা সংখ্যাবাচক ভ্যারিয়বল, এবং কোয়ালিটেটিভ (qualitative) বা ক্যাটেগরিক্যাল (categorical) ভ্যারিয়েবল।

যে ভ্যারিয়েবল শুধু নাম্বার বা সংখ্যা মান গ্রহণ করে তাদেরকে কোয়ান্টিটেটিভ ভ্যারিয়েবল বলে। কোয়ান্টিটেটিভ ভ্যারিয়েবলের উপর সাধারণ গাণিতিক অপারেশন যেমন যোগ, বিয়োগ, গুন, ভাগ- এসব প্রয়োগ করা যায়।

যে ভ্যারিয়েবল শুধু নাম জাতীয় মান গ্রহণ করে তাদের কোয়ালিটেটিভ ভ্যারিয়েবল বলে। যেমন, আজ কী বার এটি একটি কোয়ালিটেটিভ ভ্যারিয়েবল। স্বভাবতই কোয়ালিটেটিভ ভ্যারিয়েলের উপর গাণিতিক অপারেশন করা যায়না। কেননা আজ এবং কাল এই দুই দিনকে আমরা যোগ করবে পারব না।

অনেক সময় কোয়ালিটেটিভ ভ্যারিয়েবলের মান সংখ্যা দিয়ে প্রকাশ করা হয়। যেমন সপ্তাদের দিনগুলোকে Saturday, Sunday এসব না বলে আমরা 1, 2, 3, 4, 5, 6, 7 এসব দিয়ে প্রকাশ করতে পারতাম। সেক্ষেত্রে কম্পিউটার এই ভ্যারিয়েবলতে হয়তো সংখ্যা হিসেবে স্টোর করত। কিন্তু তার মানে এই নয় যে আমরা গাণিতিক অপারেশন করতে পারব। মূল কথা, কোয়ালিটেটিভ ভ্যারিয়েবল কেবলমাত্র বর্ণানামূলক বৈশিষ্ট্য নির্দেশ করে।

আরোও কিছু উদাহরণ দেখা যাক।

ভ্যারিয়েবলের ধরণ	উদাহরণ	ডেটা
কোয়ান্টিটেটিভ	ছাত্রের বয়স (বছর)	২৫, ২৩, ২০
-	ব্যাংকে টাকার পরিমাণ	১০০০০০, ৫৬০০০, ৩৭০০
-	দৈনিক গড় তাপমাত্রা	২৩, ৩৩, ৪১
-	ট্রাফিক জ্যামে দৈনিক ব্যয় (মিনিট)	১২০, ২০০, ১৮৭
কোয়ালিটেটিভ/ক্যাটেগরিক্যাল	পরীক্ষার ফল	পাশ, ফেইল
-	অর্থনৈতিক ক্ল্যাস	উচ্চ বিত্ত, মধ্যবিত্ত, নিম্নবিত্ত
-	হালের জনপ্রিয় সাবজেক্ট	স্ট্যাটিসটিক্স, ডেটা সায়েন্স

ভ্যারিয়েবলের ধরণ	উদাহরণ	ডেটা
-	ডেটা সায়েন্স প্রোগ্রামিং ল্যাঙ্গুয়েজ	পাইথন, R, জুলিয়া

লক্ষ্য করুন, কোয়ান্টিটেটিভ ভ্যারিয়েবলগুলি কিন্তু সবই একই রকম। কিন্তু কোয়ালিটেটিভ বা ক্যাটেগরিক্যাল ভ্যারিয়েবলগুলোর মধ্যে একটি আছে যেটি অন্যদের থেকে একটু আলাদা। পরীক্ষার ফল, জনপ্রিয় সাবজেক্ট, প্রোগ্রামিং ল্যাঙ্গুয়েজ- - এগুলোর কোন ন্যাচারাল অর্ডার নেই। অর্থাৎ কোনটি বড় বা উচু কিংবা কোনটি ছোট বা নিচু এরকম করে সাজানো যায় না। তবে অর্থনৈতিক ক্ল্যাস- - এই ভ্যারিয়েবলের মানগুলোকে ক্রমাকারে সাজানো যায়- - বড় থেকে ছোট বা ছোট থেকে বড়। ডেটা এনালাইজ করার সময় ও স্ট্যাটিসটিক্যাল প্রেডিক্টিভ মডেল ডেভলপ করার সময় এই বৈশিষ্ট্যটি কাজে লাগে।

২.৩ এক্সারসাইজ

এবার একটি টেবিল তৈরী করুন যেখানে উপরের ডেটাগুলিতে যে ভ্যারিয়বলগুলো আছে সেগুলো কোনটি কী ধরনের ভ্যারিয়েবল সেটি লিখুন।

অধ্যায় ৩

ডেটা ডিস্ট্রিবিউশন

অধ্যায়ের নামটি কঠিন হয়ে গেল।

পাঠক, ধারণা করছি আপনি এই শব্দের সাথে পরিচিত নন। এই অধ্যায়ে পরিসংখ্যান তথা ডেটা এনালিসিসের অন্যতম গুরুত্বপূর্ণ ধারণাটি আমরা জানার চেষ্টা করব। ডিস্ট্রিবিউশন যদি বোঝা যায় তাহলে প্রবাবিলিটি ডিস্ট্রিবিউশন কী সেটিও সহজেই বোঝা যাবে। আর সে কারণেই এই অধ্যায়ে ডেটা ভিজুয়ালাইজেশনের মাধ্যমে আমরা ডেটা ডিস্ট্রিবিউশনের ব্যাপারটি রপ্ত করব।

প্রথমত এই অধ্যায়ের মূল লক্ষ্য ডেটা ভিজুয়ালাইজেশন। দ্বিতীয়ত ডেটা ভিজুয়ালাইজেশনের মাধ্যমে পরিসংখ্যানের মৌলিক কিছু কনসেপ্ট সম্পর্কে মনের মধ্যে আপনা থেকেই প্রশ্ন তৈরী করব। তারপর সেই প্রশ্নের উত্তর আমরা খোঁজার চেষ্টা করবো। ডেটাকে সংখ্যা বা টেবিলের মাধ্যমে না দেখে, চিত্র বা গ্রাফের মাধ্যমে দেখলে অনেক সময় এমন তথ্য পাওয়া যায় যা অন্যভাবে দেখলে পাওয়া যায় না। সে কারণে ডেটা ভিজুয়ালাইজেশনের প্রতি গুরুত্ব দিচ্ছি।

এই অধ্যায়ে ডেটা ভিজুয়ালাইজেশনের জন্য যে কম্পিউটার কোড ব্যবহার করা হচ্ছে সেগুলো আপনি বুঝতে না পারলেও কোন সমস্যা নেই। যারা R সফটওয়্যারের সাথে পরিচিতি তারা ইচ্ছে করলে কোডগুলো আপনাদের কম্পিউটারে চালিয়ে দেখতে পারেন। নিজের অনুশীলনের জন্য কোডগুলো আশা করি কাজে দেবে।

প্রথমে আমাদের ডেটা খুঁজে বের করতে হবে। বাংলাদেশ সরকার অনেক ডেটা এখন অনলাইনে পাবলিশ করে। সেখান থেকে কোন একটি ডেটা আমরা নিয়ে ব্যাবহার করেতে পারি।

৩.১ ডিস্ট্রিবিউশন কী?

ডিস্ট্রিবিউশন ইংরেজীতে distribution, আর বাংলায় বিন্যাস। যার অর্থ ছড়িয়ে থাকা। অর্থাৎ কোন কিছু যেভাবে ছড়িয়ে আছে সেটিকে তার বিন্যাস বলে। আমরা বিন্যাস শব্দটি ব্যবহার করব না; আমরা ব্যবহার করব ডিস্ট্রিবিউশন শব্দটি। আমরা চাই ডিস্ট্রিবিউশন শুনলেই আমাদের চোখে যেন ভেসে ওঠে ডিস্ট্রিবিউশন বলতে কী বোঝায়। যেমন

- বাংলাদেশের মানুষেরা নানা বিভাগে কীভাবে ডিস্ট্রিবিউটেড বা ছড়িয়ে আছে সেটিকে বলব পপুলেশন ডিস্ট্রিবিউশন
- সুন্দরবনে বাঘগুলো বনের কোথায় কোথায় কীভাবে ছড়িয়ে আছে, সেটিকে বলব বাঘের ডিস্ট্রিবিউশন
- আপনার পড়ার টেবিলে বইগুলো যেভাবে ছড়িয়ে আছে সেটিকে আমরা বলতে পারি বইগুলোর ডিস্ট্রিবিউশন

এ তো গেল অতি পরিচিত কিছু উদাহরণ। উল্লিখিত উদাহরণের সবগুলি যে টেকনিক্যালি নিখুঁত তা নয়। কিন্তু ডিস্ট্রিবিউশন বলতে কী বোঝায় সেটি বোঝাটাই আসল।

ডিস্ট্রিবিউশনের প্রাথমিক ধারণা থেকে আমাদের মনের মধ্যে নানা রকম প্রশ্ন জাগতে পারে। যেমন, ছড়িয়ে থাকা বলতে আমরা কী বোঝাচ্ছি? বাংলাদেশের পপুলেশন ডিস্ট্রিবিউশন বলতে আমরা আসলে কী বোঝাচ্ছি?

এর উত্তর হতে পারে নানা রকম। যেমন

- বাংলাদেশের জনসংখ্যার কতভাগ পুরুষ আর কতভাড় স্ত্রী সেটি এই পপুলেশনের জেন্ডার ভিত্তিক ডিস্ট্রিবিউটশন (sex distribution)
- জনসংখ্যার কত সংখ্যক বিভিন্ন বয়সগ্রুপে আছে সেটি পপুলেশনের বয়সভিত্তিক ডিস্ট্রিবিউটশন (age distribution)। যেমন ০-৫ বছর বয়সী জনসংখ্যা, ৫-১০ বছর বয়সী জনসংখ্যা, ইত্যাদি
- তেমনি বাংলাদেশের কোন জেলাতে কী পরিমাণ মানুষ বাস করে সেটি জনসংখ্যার স্পেশিয়াল ডিস্ট্রিবিউটশন (spatial distribution)

এখন ভাবুন তো ফেইসবুক পোস্টে লাইকের ডিস্ট্রিবিউটশন বলতে আমরা তাহলে কী বুঝব? এই ডিস্ট্রিবিউটশনও নানা রকমের হতে পারে। যেমন

- বন্ধুদের জেন্ডারের ভিত্তিতে লাইকের ডিস্ট্রিবিউটশন। ছেলে ও মেয়ে বন্ধুদের লাইক করার শতকরা হার
- ০-১০ টি লাইক পেয়েছে এমন পোস্টের সংখ্যা, ১১-২০ টি লাইক পেয়েছে এমন পোস্টের সংখ্যা, এভাবে পোস্টে লাইকের ডিস্ট্রিবিউটশন

এভাবে দৈনিক ইন্টারনেট ব্যবহারের ডেটা থেকে আমরা ইন্টারনেট ব্যবহারের ডিস্ট্রিবিউটশন সম্পর্কেও কিছুটা ধারণা করতে পারব। যেমন

- সপ্তাহভিত্তিক ইন্টারনেট ডেটা ব্যবহারের ডিস্ট্রিবিউটশন
- ০ থেকে ২৫০ মেগা ডেটা কতবার ব্যবহার করেছি, ২৫০ থেকে ৫০০ মেগা কতবার ব্যবহার করেছি, এভাবে ডেটা ব্যবহারের পরিমাণকে যদি বিভিন্ন শ্রেনীতে ভাগ করি তাহলে সেটা হবে ইন্টারনেট ডেটা ব্যবহারের ডিস্ট্রিবিউটশন

৩.২ পরিসংখ্যানে ডিস্ট্রিবিউটশন কী?

উপরের উদাহরণগুলো অবশ্যই ডিস্ট্রিবিউটশনের উদাহরণ। বাস্তবজীবনের নানা ঘটনাকে আমরা এভাবে ডিস্ট্রিবি-উশনের সংজ্ঞায় ফেলতে পারি। তবে পরিসংখ্যানের ডিস্ট্রিবিউশনের সুনির্দিষ্ট অর্থ রয়েছে। উপরের উদাহরণগুলো থেকে সেটি আলাদা কিছু নয়। বরং উপরের উদাহরণগুলোর মধ্যেই আছে পরিসংখ্যান ডিস্ট্রিবিউটশন বলতে যা বোঝায়।

পরিসংখ্যানে ডিস্ট্রিবিউশন শব্দটি ব্যবহৃত হয় ভ্যারিয়েবলের সাথে যুক্ত করে। অর্থাৎ ডিস্ট্রিবিউটশন বলতে ভ্যারিয়েবলের ডিস্ট্রিবিউটশন বোঝায়। আরো সুনির্দিষ্ট করে বললে র‍্যান্ডম ভ্যারিয়েবলের ডিস্ট্রিবিউটশন বে-োঝায়। তবে র‍্যান্ডম ভ্যারিয়েবল ও তার ডিস্ট্রিবিউটশন আমরা পরে জানব। আপাতত ভ্যারিয়েবলের ডিস্ট্রিবি-উটশন বুঝতে চেষ্টা করি।

৩.২.১ ভ্যারিয়েবলের ডিস্ট্রিবিউটশন

মনে করার চেষ্টা করুন ভ্যারিয়েবল কী?

ভ্যারিয়েবল হলো ডেটার মধ্যে যে কলাম গুলো থাকে সেগুলো। ভ্যারিয়েবলের মান হতে পারে সংখ্যা (যেমন, ১, ২, ইত্যাদি) বা মান (যেমন, নাম, সময়)। এই মানগুলো ডেটার প্রত্যেকটি সারির জন্য একই হয় না। অর্থাৎ ভ্যারিয়েবল একেক সময় একেক মান গ্রহণ করতে পারে। একাধিক সারিতে একই মানও হতে পারে। কিন্তু এমন হয়না যে ভ্যারিয়েবলের সবগুলো মান একই। সেক্ষেত্রে সেটি হবে ধ্রুবক বা ফিক্সড। কোন একটি ভ্যারিয়েবল কী মান গ্রহণ করবে সেটি সেই ভ্যারিয়েবলের ধরণ ও ডিস্ট্রিবিউশনের উপর নির্ভর করে। প্রশ্ন হচ্ছে ভ্যারিয়েবলের ডিস্ট্রিবিউটশন বলতে কী বোঝায়?

লক্ষ্য করুন, ভ্যারিয়েবলের ডিস্ট্রিবিউটশন ঐ ভ্যারিয়েবল কী মান গ্রহণ করতে পারে তার সাথে সম্পর্কযুক্ত। অর্থাৎ ভ্যারিয়েবলের মানগুলো কী কী হতে পারে এবং সে মান গুলো কীভাবে বিন্যস্ত হয় তাকে সেই ভ্যারিয়েবলের ডিস্ট্রিবিউটশন বলে। নীচের টেবিলে উদাহরণ দিয়ে ব্যপারটি সহজ করে দেখানোর চেষ্টা করি।

ভ্যারিয়েবল	বাস্তবে কাজের প্রশ্ন যার উত্তর জানতে চাই	ভ্যারিয়েবলের সম্ভাব্য মান
ফেইসবুক লাইক	লাইক যে দিয়েছে সে ছেলে না মেয়ে?	ছেলে, মেয়ে
-	পোস্ট করার প্রথম ঘন্টায় কয়টি লাইক পড়েছে? প্রতি ঘন্টায় কয়টি লাইক পড়ছে?	০, ১, ২, বার
-	১০ - ২০ টি লাইক পড়েছে কতবার? ৫০ এর কম লাইক পড়েছে কতবার? ১০০টির বেশী লাইক পড়েছে কতবার?	০, ১, ২, বার
ইন্টারনেট ডেটা ব্যবহার	সপ্তাহে মোট ডেটা ব্যবহারের পরিমাণ	শূন্য বা শূন্য থেকে বড় যেকোন সংখ্যা
-	দিনে ১ গিগাবাইট ডেটা ব্যবহার হয়েছে কতবার?	০, ১, ২, বার

সারনী ৩.২: লিংগভেদে ইন্টারনেট ব্যবহারকারির ডিস্ট্রিবিউশন

	Frequency	Percent
Female	124	51.03
Male	119	48.97
Total	243	100.00

সারনী ৩.৩: লিংগভেদে ইন্টারনেট ব্যবহারের গড় (মেগাবাইট)

sex	Avg_Usage
Female	159.8
Male	154.4

ভ্যারিয়েবল	বাস্তবে কাজের প্রশ্ন যার উত্তর জানতে চাই	ভ্যারিয়েবলের সম্ভাব্য মান
মিরপুর-মতিঝিল কমিউট টাইম	সকালে রওনা দিলে দুই ঘন্টার মধ্যে মতিঝিল পৌঁছেছে এমন কতদিন হয়েছে। পরে এর সম্ভাবনা আমরা জানতে চাইব।	০, ১, ২, □ বার

চলুন তাহলে ফেইসবুক লাইক ডেটা, ইন্টারনেট ডেটা ব্যবহারের পরিমাণ, কমিউট সময়□ এসব ভ্যারিয়েবলের ডিস্ট্রিবিউটশন কেমন। ভ্যারিয়েবলের ডিস্ট্রিবিউটশন দেখতে হলে আমাদেরকে চিত্রের সাহায্য নিতে হবে। স্ট্যাটিসটিক্যাল সফটওয়্যারের সাহায্যে খুব সহজেই আমরা ডেটা ডিস্ট্রিবিউটশন দেখতে পারব।

৩.৩ ফ্রিকোয়েন্সি ডিস্ট্রিবিউটশন

ভ্যারিয়েবলের মানগুলো কোনটি কতবার ঘটছে সেটিকে একটি টেবিলের মাধ্যমে প্রকাশ করলে তাকে ফ্রিকোয়েন্সি ডিস্ট্রিবিউটশন বলে। ফ্রিকোয়েন্সি ডিস্ট্রিবিউটশন সাধারণত ক্যাটেগরিক্যাল ডেটার জন্য করা হয়। তবে কোয়ান্টিটেটিভ ডেটার জন্যও করা যায়। কিন্তু সেটি কদাচিত করা হয়ে থাকে কেননা কোয়ান্টিটেটিভ ভ্যারিয়েবলের ফ্রিকোয়েন্সি টেবিল খুব কাজের নয়। ডেটার আকার ছোট হলে কিংবা ভ্যারিয়েবলের মান অল্পসংখ্যক হলে সেক্ষেত্রে ফ্রিকোয়েন্সি টেবিল কাজে দিতে পারে। যেমন লিংগভেদে ইন্টারনেট ব্যবহারকারির ডিস্ট্রিবিউটশন টেবিল ৩.২ তে দেয়া হল।

এই ডিস্ট্রিবিউটশন থেকে আমরা দেখতে পাই ইন্টারনেট ব্যবহারকারির 51.03% মহিলা এবং বাকীরা পুরুষ। তেমনি লিংগভেদে গড় ইন্টারনেট ব্যবহারের পরিমাণও আমরা দেখতে পারি।

৩.৪ হিস্টোগ্রাম

ডেটার ডিস্ট্রিবিউটশন দেখতে দুই ধরনের চিত্র ব্যবহার করা যায়। ক্যাটেগরিক্যাল ডেটার জন্য ফ্রিকোয়েন্সি টেবিল আর কোয়ান্টিটেটিভ ডেটার জন্য হিস্টোগ্রাম। পরিসংখ্যানে ফ্রোকোয়েন্সি বলতে কোন ভ্যারিয়েবলের মান কতবার ঘটছে সেটি বোঝায়। যেমন, ১০ জন ছাত্র-ছাত্রীর মধ্যে ৪ জন ছেলে আর ৬ জন মেয়ে। এই ৪ হল ছাত্রের ফ্রিকোয়েন্সি আর ৬ হলো ছাত্রীর ফ্রিকোয়েন্সি। অর্থাৎ ডেটার মধ্যে ছাত্র ও ছাত্রী কতবার আছে সেটিকেই ফ্রিকোয়েন্সি বলে।

আর কোয়ান্টিটেটিভ ডেটার জন্য আঁকা হয় হিস্টোগ্রাম। হিস্টোগ্রাম কীভাবে হাতে কলমে করা হয় সেটি এই বইয়ের উদ্দেশ্যের বাইরে। সে জন্য পরিসংখ্যানের যেকোন পাঠপুস্তক দেখে নিতে হবে। এখানে হিস্টোগ্রামের একটা ওভারভিউ দেয়া হবে এবং মূলত হিস্টোগ্রাম ব্যবহার করে কীভাবে ডেটার ডিস্ট্রিবিউটশন সম্পর্কে জানা যাবে সেটি দেখানো উদ্দেশ্য।

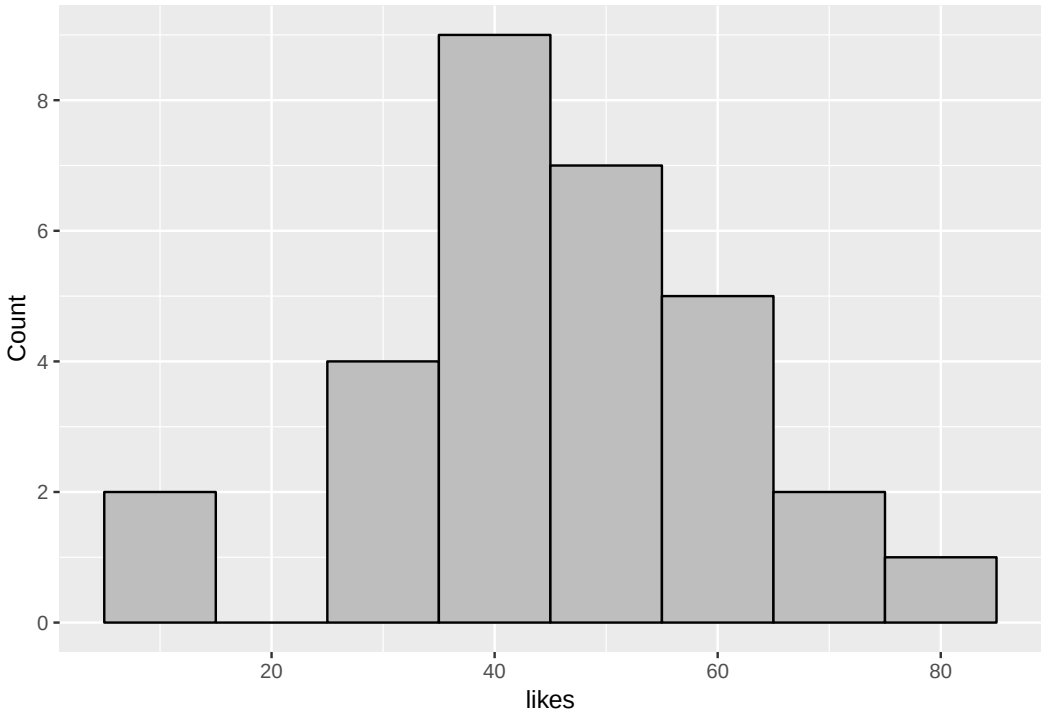
হিস্টোগ্রাম বানানোর জন্য প্রথমে ডেটাকে আমরা কয়েকটি ঝুড়ি বা বাকেট (bucket) বা বিন (bin)-এ ভাগ করব। এরপর প্রতিটি বিনে কয়টি ডেটা আছে সেটি গুনব। ফেইসবুক লাইকের ডেটার দিতে তাকালে আমরা দেখব শূন্যের নীচে লাইক সংখ্যা হতে পারে না। আর আমাদের ডেটায় সর্বোচ্চ লাইকের সংখ্যা 77। তাহলে শূন্য থেকে ৮০ পর্যন্ত যদি ১০টি বিন বানাই তাহলে সেগুলো হবে ০-১০, ১০-২০, ২০-৩০, ৩০-৪০, ৪০-৫০, ৫০-৬০, ৬০-৭০, ৭০-৮০।

যদিও ফ্রিকোয়েন্সি টেবিল ডেটা সামারাইজ করার জন্য

```
summary(fblikes$likes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   41.00   44.50   45.53   55.50   77.00
```

```
library(ggplot2)
ggplot(fblikes, aes(x = likes)) +
  geom_histogram(color = "black", fill = "gray", binwidth = 10) +
  scale_y_continuous("Count", c(seq(0,10,2)))
```



৩.৫ কমিউট টাইম (মিরপুর টু মতিঝিল)

কমিউট টাইম

```
## Parsed with column specification:
## cols(
##   dates = col_date(format = ""),
##   day_name = col_character(),
##   bus = col_double(),
##   uber = col_double(),
##   pathao = col_double(),
##   time_of_day = col_character()
## )
```

```
names(commute_morning)
```

```
## [1] "dates"       "day_name"    "bus"         "uber"        "pathao"
## [6] "time_of_day"
```

অধ্যায় ৪

টাইডি ডেটা

Some significant applications are demonstrated in this chapter.

৪.১ Example one

৪.২ Example two

অধ্যায় ৫

রিগ্রেশন

এর আগে আমরা ইউনিভ্যারিয়েট ডেটার উদাহরণ দেখেছি। বাস্তব জীবনে একক ভ্যারিয়েবলগুলো একটির সাথে আরেকটি নানা ভাবে সংযুক্ত থাকে। অর্থাৎ ভ্যারিয়েবলগুলো একে অপরের সাথে সম্পর্কযুক্ত। কখনও কখনও সম্পর্ক গুলো দূর্বল হয় আবার কখনও অনেক শক্ত হয়।

দুটি নিউমেরিক্যাল ভ্যারিয়েবলের মধ্যেকার সরলরৈখিক সম্পর্ক পরিমাপ করার জন্য কোরিলেশন কোএফিশিয়েন্ট ব্যবহার করা হয়। কোরিলেশন কোএফিশিয়েন্টকে আমরা ρ দিয়ে চিহ্নিত করেছিলাম। ρ এর মান -১ থেকে +১ এর মধ্যে থাকতে পারে। মান যদি -১ এর কাছাকাছি হয় (যেমন -০.৯০) তাহলে আমরা দৃঢ় (strong) নেগেটিভ কোরিলেশন বলি। আর এর মান যদি +১ (যেমন +০.৯০) এর কাছাকাছি হয় তাহলে সেটিকে দৃঢ় পজিটিভ কোরিলেশন বলি। এর মান ০ এর কাছাকাছি হলে দূর্বল পজিটিভ (যেমন -০.১৫) বা দূর্বল নেগেটিভ (যেমন +০.১৫) কোরিলেশন বলি। এখানে ০.৯০ বা ০.১৫ মানগুলো উদাহরণ হিসেবে নেয়া হয়েছে। কোরিলেশনের কোন মানটি বড় বা কোন মানটি ছোট তার কোন নির্ধারিত তালিকা নেই। এটি ব্যবহারকারির উপর অনেকাংশে নির্ভর করে।