

# ফলিত পরিসংখ্যান ও ডেটা সায়েন্স

এনায়েতুর রহীম

2018-09-01



# সূচীপত্র

১	ভূমিকা	৫
১.১	পরিকল্পনা . . . . .	৫
১.২	ভূমিকা . . . . .	৫
১.৩	ডেটা কী ও দেখতে কেমন? . . . . .	৬
১.৪	ডেটা ভিজুয়ালাইজেশন . . . . .	৬
১.৫	লোকেশন ও ভ্যারিয়েশনের পরিমাপ . . . . .	৬
১.৬	ভ্যারিয়েবলের মধ্যে সম্পর্ক . . . . .	৬
১.৭	প্রবাবিলিটি সম্পর্কে ধারণা . . . . .	৬
১.৮	. . . . .	৬
২	ডেটা সায়েন্স	৯
২.১	ডেটা সায়েন্স কী? . . . . .	৯
২.২	ডেটা সায়েন্স কোথায় ব্যবহৃত হচ্ছে? . . . . .	১০
২.৩	মেশিন লারনিং . . . . .	১১
২.৪	ডেটা সায়েন্টিস্টদের কাজের ধরণ . . . . .	১১
২.৫	ডেটা সায়েন্স- আপনার উপযুক্ত জব কোনটি? . . . . .	১১
২.৬	Data Scientist at GE . . . . .	১২
২.৭	সোর্সিং . . . . .	১৩
২.৮	ইটিএল . . . . .	১৪
২.৯	পরিসংখ্যানে দক্ষতা অর্জন যে কারণে প্রয়োজন . . . . .	১৪
২.১০	বইটি যাদের জন্য . . . . .	১৬
২.১১	বইটি যেভাবে পড়তে হবে . . . . .	১৬
২.১২	R পরিচিতি . . . . .	১৬
২.১৩	বাংলা কী বোর্ড পরীক্ষা করা হচ্ছে . . . . .	১৬
৩	ডেটা ও ভ্যারিয়েবল	১৯
৩.১	ডেটার উদাহরণ . . . . .	১৯
৩.২	ভ্যারিয়েবল ও প্রকারভেদ . . . . .	২০

৪	ডেটা ভিজুয়ালাইজেশন	২৩
৫	হাইপোথিসিস টেস্টিং	২৫
৫.১	Example one . . . . .	২৫
৫.২	Example two . . . . .	২৫
৬	সামারি বা সারাংশ	২৭

# অধ্যায় ১

## ভূমিকা

### ১.১ পরিকল্পনা

বইটিতে যেসব অধ্যায় থাকবে বলে ঠিক করেছি তার একটি তালিকা দিচ্ছি। বই লেখাকালীন সময়ে এটি পরিবর্তিত ও পরিবর্ধিত হতে পারে।

### ১.২ ভূমিকা

এখানে পরিসংখ্যানের একটি সাধারণ ধারণা দেয়া হবে। সেই সাথে ডেটা সায়েন্স নিয়ে দুচারটা কথা বলা হবে। উদ্দেশ্য হলো পাঠককে আগ্রহী করে তোলা। ভালো হবে কয়েকটি উদাহরণ দিয়ে বোঝাতে পারলে। ছবি সহ উদাহরণ দিলে আগ্রহী সৃষ্টি হবে।

১.৩ ডেটা কী ও দেখতে কেমন?

১.৪ ডেটা ভিজুয়ালাইজেশন

১.৪.১ ডেটার শেইপ নিয়ে আলোচনা

১.৫ লোকেশন ও ভ্যারিয়েশনের পরিমাপ

১.৫.১ মিন, মিডিয়ান, মোড

১.৫.২ রেইঞ্জ, ভ্যারিয়্যান্স, স্ট্যান্ডার্ড ডিভিয়েশন

১.৬ ভ্যারিয়েবলের মধ্যে সম্পর্ক

১.৬.১ বাইভ্যারিয়েট সম্পর্ক

১.৬.২ মাল্টিভ্যারিয়েট সম্পর্ক

১.৬.৩ রিগ্রেশনের ধারণা

১.৭ প্রবাবিলিটি সম্পর্কে ধারণা

১.৭.১ সেট থিউরি

১.৭.২ ইভেন্ট

১.৮

আমার দেশের নাম বাংলাদেশ। ইস্টিশন।

আমার সোনার বাংলাদেশ।

বাংলায় ফলিত পরিসংখ্যান বই লিখতে বসে অসম্ভব ভালো লাগছে। এই কারণে যে পরিসংখ্যানের ছাত্র হয়ে এই ফিল্ডে সামান্য হলেও খানিকটা অবদান রাখতে আল্লাহ আমাকে তৌফিক দিয়েছেন।

এটি ইটালিক টেক্সট

This is a sample book written in Markdown. You can use anything that Pandoc's Markdown supports, e.g., a math equation  $a^2 + b^2 = c^2$ .

For now, you have to install the development versions of bookdown from Github:

```
devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading #.

To compile this example to PDF, you need to install XeLaTeX.





## অধ্যায় ২

### ডেটা সায়েন্স

ডেটা সায়েন্স এখন হট টপিক। ডেটা সায়েন্টিস্ট-দের জব মার্কেটে অনেক ডিম্যান্ড। বিদেশের মতো বাংলাদেশে ডেটা সায়েন্সের ব্যবহার এখনো সেরকমভাবে শুরু না হলেও ভেতরে ভেতরে তার প্রস্তুতি চলছে। কিছু কিছু ক্ষেত্রে বেশ অগ্রগতিও হয়েছে। মার্কেটের দিকে তাকালেই আমরা কিছুটা আন্দাজ করতে পারি। বিশেষ করে তরুণ উদ্যোক্তা, ইন্টারনেট-ভিত্তিক ব্যবসা প্রসারের চিন্তাভাবনা, সরকারের সদিচ্ছা এবং তরুণ মন্ত্রী ও উপদেষ্টাদের ডিজিটাল বাংলাদেশ গড়ার প্রচেষ্টায় পজিটিভ কাজ করা এবং সর্বোপরি ইন্টারনেট ব্যবহারকারীদের এক্সপোনেনশিয়াল বৃদ্ধি অচিরেই আমাদের ব্যবসা, বাণিজ্য, মার্কেটিংকে এমনভাবে বদলে দেবে যা আমরা এই মুহূর্তে অনেকেই বিশ্বাস করতে পারছি না।

যে কারণে নিকট ভবিষ্যতে আমাদের অনেক ডেটা সায়েন্টিস্ট দরকার হবে। দেশের চাহিদা মেটাতে আমাদের নিজেদের মধ্যে থেকেই সেই স্থান পূরণ করতে হবে। আমার কাছে অনেকেই প্রশ্ন করে কিভাবে ডেটা সায়েন্টিস্ট হওয়া যায়। আমি কথোপকথনের মাধ্যমে পুরো প্রসেসটার একটা সারাংশ দেয়ার চেষ্টা করেছি। বিস্তারিত জানতে পুরো লেখাটি পড়ার আহবান জানাচ্ছি।

কিছুদিন আগেও ডেটা সায়েন্স কী বা ডেটা সায়েন্টিস্ট কারা সেটি নিয়ে স্বচ্ছ কোন ধারণা অনেকেরই ছিল না। এখনো ততটা পরিষ্কার না হলেও ডেটা সায়েন্স কর্মক্ষেত্রে এখন নতুন কোন শব্দ নয়। এর নানাবিধ ব্যবহার থাকায় ডেটা সায়েন্স এখনো সে অর্থে সংজ্ঞায়িত করা সম্ভব হয়নি।

#### ২.১ ডেটা সায়েন্স কী?

ডেটা সায়েন্স নতুন এবং দ্রুত জনপ্রিয় হচ্ছে এমন একটি ফিল্ড। অনেকেই ইদানীং ডেটা সায়েন্সের প্রতি আগ্রহী হয়ে উঠছে। অথচ তাদের অনেকেই পরিষ্কারভাবে জানে না ডেটা সায়েন্স বলতে কী বোঝায়।

ডেটা সায়েন্স ইজ লুজলি ডিফাইন্ড। অর্থাৎ ডেটা সায়েন্স বলতে অনেক কিছুই বোঝানো সম্ভব এবং অনেক কিছুই বোঝানো হচ্ছে। কোন্ ডিসিপ্লিনের মানুষ শব্দটি ব্যবহার করছে তার উপর নির্ভর করে তারা ডেটা সায়েন্স বলতে কী

বোঝাচ্ছে। এর কারণ হলো ডেটা নিয়ে যে বা যারা কাজ করে তারাই নিজেদেরকে ডেটা সায়েন্টিস্ট হিসেবে দাবী করতে পারে।

সহজ করে বলতে গেলে ডেটা সায়েন্স এমন একটি শাখা যেখানে নানা বিষয়ের লোকজন একসাথে ডেটা নিয়ে কাজ করে সেই ডেটাকে অর্থবহ করে ব্যবহারোপযোগী করে তোলে যা থেকে সিদ্ধান্ত গ্রহণ করা সম্ভব হয়। অর্থাৎ কোন প্রতিষ্ঠানের কাছে যে ডেটা আছে সেই ডেটা প্রতিষ্ঠানটির কী কাজে লাগবে সেভাবে ডেটাকে কাজে লাগানোর বিজ্ঞানকেই ডেটা সায়েন্স বলা যাবে।

## ২.২ ডেটা সায়েন্স কোথায় ব্যবহৃত হচ্ছে?

অতি পরিচিত উদাহরণ হলো গুগল। আপনি গুগলের সার্ভিসে লগিন থাকা অবস্থায় সার্চ করেন [cricket] তাহলে যা পাবেন আপনি গুগলে লগিন না করে যদি একই কী-ওয়ার্ড সার্চ করেন দেখবেন ফলাফল একটু হলেও ভিন্ন হবে। কেন এমনটা হয় সেটা ভেবে দেখেছেন কখনো? কারণ হলো গুগল আপনার পছন্দ অপছন্দ অনেক কিছুই জানে। আর সে মোতাবেক আপনার জন্য personalized search result তৈরী করে আপনার সামনে পেশ করে। গুগল এমনটি করে থাকে আর্টিফিশিয়াল ইন্টেলিজেন্সের মাধ্যমে যেটি ডেটা সায়েন্সের একটি অংশ বলা যায়।

আপনি কী করেন, কোথায় ঘুরতে যান, কোথায় খান, কোন দোকানে শপিং করেন, কী ধরনের শপিং করেন সেসবের অনেককিছুই এখন ট্রাক করা সম্ভব এবং করা হচ্ছেও। লোকেশনের কথাই ধরা যাক। আপনার হাতের স্মার্ট ফোনটি আপনার গতিবিধি সবকিছুই রেকর্ড করছে। আপনি হয়তো জানেন না অথবা জেনেও কিছু বাড়তি সুবিধা দিয়ে গুগল আপনার কাছ থেকে এই তথ্যগুলো যোগাড় করে নিচ্ছে। এতে গুগলের ব্যবসায়িক উদ্দেশ্য তো অবশ্যই আছে সেই সাথে আপনাকেও অনেক কিছু সহজে করে দিচ্ছে। এসব তথ্য নিয়ে গুগলের মেশিন আপনার সম্পর্কে লার্ণ করছে বা শিখছে। আপনাকে চিনছে।

এটা গেল একটা উদাহরণ। ডেটা সায়েন্সে এখন বিগ ডেটার ব্যবহার শুরু হয়েছে। বিগ ডেটা হলো অনেক বড় ডেটা। যত বড় হলে আপনার কম্পিউটারে সেই ডেটা এনালাইজ করা যাবে না সেরকম ডেটাকেই বিগ ডেটা বলা হচ্ছে। অন্যভাবে বলা যায় সাধারণ কনজিউমার লেভেলের কম্পিউটার দিয়ে যে ডেটা এনালাইজ করা যায়না সেই ডেটা হলো বিগ ডেটা।

এরকম ডেটা এখন হরহামেশা তৈরী হচ্ছে। কারণ তথ্যের প্রবাহ বেড়েছে আর সেই সাথে কমেছে কম্পিউটার স্টোরেজ-এর মূল্য। ফলে সবাই এখন ব্যাপক হারে ডেটা জেনারেট করছে (যেমন ওয়েব সাইট ক্লিক, সাইট ভিজিট, টুইট করা কিংবা ফেইসবুকে স্ট্যাটাস দেয়া, ইত্যাদি) তেমনি সেই ডেটাকে অল্প খরচে এখন সংরক্ষণ করাও সহজ হয়েছে।

তবে এর ব্যতিক্রমও আছে। যেমন প্লেনের ফ্লাইট রেকর্ড যা প্রতি সেকেন্ডে প্লেনের অবস্থান, দ্রুতি, বাতাদের প্রেশার সহ হাজারো প্যারামিটারের মান রেকর্ড করতে পারে যা থেকে ফ্লাইটকে আরো নিরাপদ করা সম্ভব বলে বলা হচ্ছে। অথচ চাইলেই এরকম ডেটা রিয়েল টাইমে সংগ্রহ করা সহজ নয়। কারণ স্যাটেলাইটের মাধ্যমে গিগাবাইট ডেটা ট্রান্সফার করা ব্যাপক ব্যয়বহুল। আপনার মোবাইলের ডেটার খরচের সাথে তুলনা করলেই সেটা অনুমান করতে পারবেন।

## ২.৩ মেশিন লারনিং

ডেটা সায়েন্সের একটি দিক হলো মেশিন লারনিং। মেশিন লারনিং বাস্তবায়ন করা খুব সহজ কাজ না হলেও ব্যাপারটা বেশ সহজ। একটি শিশুর কথা ধরা যাক। শিশুটি কিভাবে ভাষা শেখে সেটি ভেবে দেখুন। যদি বলা হয় শিশু নয়, একটি যন্ত্রকে ভাষা শেখাতে হবে তাহলে আপনি কিভাবে সেটি বাস্তবায়ন করবেন যাতে সে আপনার ভাষা বুঝতে পারে? কঠিন কাজ। কিন্তু এটিই মেশিন লারনিং। মেশিন যেভাবে শেখে সেটিই মেশিন লারনিং।

উদাহরণ দেই। হাসপাতালে অপারেশনের পর ICU তে যেসব রোগী যায় তাদের (ধরি) আনুমানিক ১০ শতাংশ মারা যায়। ধরুন আপনার হাসপাতালে সেরকম এক লক্ষ্য রোগীর তথ্য আছে। রোগীর বয়স, রোগের ধরন, অন্যান্য ক্লিনিকাল তথ্য। এই ডেটা থেকে লার্ণ করে আমরা এমন একটি স্ট্যাটিস্টিক্যাল মডেল কি তৈরী করতে পারবো যেটি নতুন সেরকম একজন রোগীর মারা যাওয়ার সম্ভাবনা বের করতে পারবে? এমনটি যদি করা যায় তাহলে আমরা আগে থেকেই সে রোগীর প্রতি নিবিড় মনোযোগ দিতে পারবো বা ব্যবস্থা নিতে পারবো। এখানে মেশিন লারনিং ব্যবহার করে ফিচার বা ফ্যাক্টরগুলোকে বের করা হচ্ছে যে ফ্যাক্টরগুলো রোগীর মারা যাওয়ার সাথে সম্পর্কিত।

মেশিন লারনিং এর জনপ্রিয় উদাহরণ হলো গুগল নাউ বা Ok Google এবং এ্যাপল-এর সিরি এসিস্ট্যান্ট। তারা আপনার ভয়েস চিনে আপনার সাথে আলাপ চালিয়ে যেতে পারে এবং ছোটখাট প্রশ্নের উত্তরও দিতে পারে।

আসছে দিনে মেশিন লারনিং-এর আরেকটি উপহার আমরা পেতে যাচ্ছি যেটি হলো স্বচালিত গাড়ি।

## ২.৪ ডেটা সায়েন্টিস্টদের কাজের ধরণ

### ২.৫ ডেটা সায়েন্স- আপনার উপযুক্ত জব কোনটি?

এটি খুবই গুরুত্বপূর্ণ একটি টপিক। কেননা জব মার্কেটে নামার আগেই জানতে হবে কোন ধরনের জবের সাথে আপনার স্কিলসেট মেলে এবং কোন ধরনের জবের ফাংশন কী।

হার্ভার্ড বিজনেস রিভিউ ২০১২ সনে লিখেছিল একবিংশ শতাব্দীর সেক্সিয়েস্ট জব হলো ডেটা সায়েন্স। এর সত্যতা প্রমাণের দরকার নেই; চাকরি বাজারে তাকালেই এর সত্যতা মেলে। আগের লেখায় বলেছিলাম ডেটা সায়েন্স এখন হট টপিক। তেমনি হট এর চাকুরির বাজার। আর বাজারের চাহিদা মেটাতে আমেরিকা এবং কানাডার অনেক বিশ্ববিদ্যালয়ে সম্প্রতি ডেটা সায়েন্স প্রোগ্রাম চালু করেছে। ধারণা করা যায় এই অনেকদিন অব্যাহত থাকবে যতদিন না বাজারে ডেটা সায়েন্টিস্টদের চাহিদার সাথে যোগানের সমন্বয় হয়।

বাংলাদেশে এখনো সেভাবে ডেটা সায়েন্স-এর প্রচার কিংবা ডেটা সায়েন্টিস্টদের প্রয়োজনীয়তা অনুভূত হয়নি। তবে অন্যান্য দেশের মতো বাংলাদেশেও ব্যাপকভাবে ডেটা ভিত্তিক সিদ্ধান্তগ্রহণের কাজ দ্রুত শুরু হয়ে যাবে। এখন একটু একটু করে ঢেউ বাংলাদেশে এসে লাগছে। আমার ধারণা আগামী ১ থেকে ২ বছরের মধ্যেই এই ঢেউ সুনামির রূপ ধারণ করবে। যার মানে হলো আমেরিকার মতো বাংলাদেশেও ডেটা ব্যবহার করে ব্যবসায়িক সিদ্ধান্তগ্রহণ করা শুরু হবে। আর তখনই ডেটা সায়েন্স জব মার্কেটে দক্ষ ডেটা সায়েন্টিস্ট ও এনালিস্টদের শুন্যতা অনুভূত হবে।

বিদেশে বিশেষ করে আমেরিকাতে গত এক বছরে ডেটা সায়েন্স ফিল্ডে চাকুরির বাজার ব্যাপক প্রসার লাভ করেছে। যার ফলে এই ফিল্ডের সাথে যে-কোনভাবে সম্পর্কযুক্ত সবাই এই মার্কেটে ঢুকতে চাইছে। ফলে প্রতিযোগিতা বাড়ছে তবে এখনো দক্ষ জনবলের অভাব রয়ে গিয়েছে। ফলে যত বেশী জবের সুযোগ আছে তত বেশী প্রার্থী পাওয়া যাচ্ছে না। কিছুদিন আগেও চাকুরি প্রার্থী এবং চাকুরি দাতা কেউই ঠিকমতো জানতো না ডেটা সায়েন্স এবং ডেটা সায়েন্টিস্টদের আসলে কোথায় কাজে লাগানো হবে। কিংবা কী ধরনের দক্ষতা সম্পন্ন লোকদের হায়ার করতে হবে। সে পরিস্থিতি এখনও খুব একটা পরিবর্তিত হয়েছে তা বলা যাবে না তবে আগের চেয়ে পরিশোধিত হয়েছে এবং চাকুরিদাতারা এখন ডেটা সায়েন্টিস্টদের কাজের পরিধি এবং কাজের ধরণ আগের চেয়ে ভালো বুঝতে পারছে এবং সে মোতাবেক দক্ষ ব্যক্তিদের হায়ার করছে। বলা যায় আগামী দিনে কাজের পরিধিগুলো আরো সুনির্দিষ্ট হবে এবং চাকুরি প্রার্থীরা জব মার্কেটে তাদের জায়গা খুঁজে পাবে।

### ২.৫.১ পরিসংখ্যানবিদরা কীভাবে এই বাজারে নিজেদের উপযোগী জব খুঁজবে?

নতুন প্রজন্মের পরিসংখ্যানবিদ বা অন্য কোয়ান্টিটেটিভ ডিসিপ্লিনের গ্রাজুয়েটরা এই বাজারে প্রবেশের জন্য উদগ্রীব। কিন্তু তাদের বেশীরভাগই ডেটা সায়েন্সে প্রাতিষ্ঠানিকভাবে প্রশিক্ষিত নয়। যে কারণে চাকুরীর বাজারে গিয়ে তারা বুঝতে পারছে না কোন পদের জন্য আবেদন করবে কিংবা কোন পদের সাথে তাদের দক্ষতা সামঞ্জস্যপূর্ণ।

কিন্তু তাদের বেশীরভাগই ডেটা সায়েন্সে প্রাতিষ্ঠানিকভাবে প্রশিক্ষিত নয়। যে কারণে চাকুরীর বাজারে গিয়ে তারা বুঝতে পারছে না কোন পদের জন্য তারা আবেদন করবে কিংবা কোন পদের সাথে তাদের দক্ষতা সামঞ্জস্যপূর্ণ। জব ডেসক্রিপশনের সাথে প্রার্থীর অভিজ্ঞতা ও দক্ষতার মিল না হলে একদিকে যেমন চাকুরী প্রাপ্তিতে দেরী হবে অন্যদিকে চাকুরী কোনভাবে পেয়ে গেলেও কর্মক্ষেত্রে গিয়ে সেই মিসম্যাচ প্রকাশিত হবে। এটা প্রার্থী এবং কোম্পানী □ দুয়ের জন্যই ব্যয়বহুল।

### ২.৫.২ ডেটা সায়েন্সে চাকুরীর বাজারে যা চাওয়া হচ্ছে

এটি অনস্বীকার্য যে সংশ্লিষ্ট সবাই এই ফিল্ডে কাজ করতে চাইছে এবং আরো অনেকেই উঁকি মেরে দেখছে যে ডেটা সায়েন্টিস্টরা আসলে কী ধরনের কাজ করে। আমি এখানে এটা হাই-লেভেল ওভারভিউ দিচ্ছি যাতে ডেটা সায়েন্স ফিল্ডে কাজের প্রবাহচিত্রটি বোঝা যায়। প্রথমেই দেখে নেয়া যাক ডেটা সায়েন্টিস্ট পদের জবে কী ধরনের অভিজ্ঞতা চাওয়া হয়।

জেনারেল ইলেক্ট্রিক তাদের ক্যালিফোর্নিয়া বেইজে ডেটা সায়েন্টিস্ট চাইছে। সেখানে যা যা চাওয়া হচ্ছে তার সারসংক্ষেপ □

## ২.৬ Data Scientist at GE

এরপর দেখা যাক ক্লিভল্যান্ড ক্লিনিকে ডেটা সায়েন্টিস্টদের জন্য কী কী চাওয়া হচ্ছে □

দেখা যাক ইন্টেল কী ধরনের ডেটা সায়েন্টিস্ট চাইছে

আপনি উপরের উদাহরণগুলোতে পুরো জব ডেসক্রিপশন নাও পড়ে থাকেন শুধু আন্ডারলাইন করা অংশটুকু পড়লেও বুঝতে পারবেন কী ধরনের দক্ষতা এরা চাইছে। পরিসংখ্যানের ছাত্র হয়ে থাকলে এর অধিকাংশই আপনার জানা-শোনার মধ্যে হবে। শুধু ডেটাবেইজ, বিগডেটা আর্কিটেকচার, হাডুপ, এইচবেইজ, মাহাউট, এসব আপনার কাছে নতুন লাগবে।

আপনি উপরের উদাহরণগুলোতে পুরো জব ডেসক্রিপশন নাও পড়ে থাকেন শুধু আন্ডারলাইন করা অংশটুকু পড়লেও বুঝতে পারবেন কী ধরনের দক্ষতা এরা চাইছে। পরিসংখ্যানের ছাত্র হয়ে থাকলে এর অধিকাংশই আপনার জানা-শোনার মধ্যে হবে।

ডেটা সায়েন্স ফিল্ডে কাজের প্রবাহচিত্র

এবারে প্রবাহচিত্রটি তুলে ধরছি যাতে করে সবাই সহজে বুঝতে পারেন। পোস্টের একদম শুরুতে আমি একটি প্রবাহচিত্র দিয়েছি। সেটিকেই বাংলায় করে দিলাম যাতে পাঠক আকৃষ্টবোধ করেন। চিত্র দেখুন।



ডেটা সায়েন্স এর জব মূলত ইটিএল (ETL □ Extract, Transform, Load) এবং এনালিটিক্স □ এই দুই ধাপে সীমাবদ্ধ। আরেকটি ধাপ চিন্তা করা যেতে পারে যেটিকে আমি তথ্য সংগ্রহ ধাপ বলেছি। এগুলো সংক্ষেপে ব্যাখ্যা করছি।

## ২.৭ সোর্সিং

প্রথম ধাপটি সাধারণত ডেটা সাইন্টিস্টদের কাছে উহ্য থাকে। যেটিকে আমরা সোর্স বলতে পারি। সোর্স পর্যায়ে আসলে তথ্য সংগ্রহ করা হয়। কোন একটি ইন্ডাস্ট্রির কথা চিন্তা করুন। তারা কোন না কোন ভাবে তথ্য সংগ্রহ করছে। এসব সিস্টেমে ফ্রন্টএন্ডে ইউজার ফ্রেন্ডলি ইন্টারফেসের মাধ্যমে তথ্য ইনপুট করা হয়। যেটি কোন এক ডেটাবেইজে গিয়ে

জমা হয়। মোটা দাগে এই অংশটি সফটওয়্যার ডেভলপারদের কাজ। তারা ইউজার ইন্টারফেইস ডিজাইন থেকে শুরু করে বেসিক ডেটা স্টোরেজ সলুশন তৈরী করে। এই ধাপে সাধারণত কোন এনালিটিক কাজ করা হয়না। হয়তো বেসিক ডেটা সামারি টাইপের কিছু করা হয়ে থাকে। সেটা নির্ভর করে সফটওয়্যারে সেরকম সুবিধাদি আছে কিনা তার উপর।

## ২.৮ ইটিএল

দ্বিতীয় ধাপটি অনেক বড়। এটিকে জব ডেসক্রিপশনে বলবে ETL। এই ধাপটিকে আইটি ধাপও বলা যায় কেননা পরিসংখ্যানবিদরা সাধারণত এই ধাপে কাজ করেন বা তারা এখন পর্যন্ত ততটা অবগত নয়। তবে পরিস্থিতি বদলাচ্ছে। এখন সবাইকেই সবকিছু সম্পর্কে মোটামুটি ধারণা রাখতে হচ্ছে।

সহজ কথায় বলতে গেলে এই ধাপে ITর লোকজন নানা সোর্স সিস্টেম থেকে ডেটাকে এনালিস্টদের কাছে পৌঁছানোর জন্য যা করা দরকার সেগুলো সম্পন্ন করে। এই ধাপে যে কাজ গুলো করা হয় সেগুলোকে সংক্ষেপে ETL বা extract, transform, load বলে। এখানে বিভিন্ন সোর্স সিস্টেমের র ডেটাকে একত্রীকৃত করে, সেগুলোকে প্রয়োজন মতো একটার সাথে একটা জুড়ে দিয়ে (merging/linking/joining), পরিশোধন, পরিবর্তন, পরিবর্তন করে ডেটা ওয়্যারহাউজে এনে জমা করে। ডেটা ওয়্যারহাউজ হলো অনেক বড় ডেটা স্টোরেজ যেখান থেকে প্রয়োজনীয় ডেটা সহজে কোয়েরি করা যায়। বলা যায় ডেটা ওয়্যারহাউজ হলো সব ডেটাবেজের ডেটাবেজ।

আমি জানি বাইরে থেকে এই প্রসেসটি সহজে বোঝা সহজ নয়। তাই বাংলাভাষীরা বুঝতে পারবে এমন একটি উদাহরণের মাধ্যমে বোঝানোর চেষ্টা করবো।

ইটিএল প্রসেসটিকে তুলনা করা যায় জমি থেকে ফসল তুলে খাবার টেবিলে পৌঁছানোর আগ পর্যন্ত যা করা হয় তার সাথে। ফসলের জমি থেকে নানা ধরনের ফসল (ধান, আলু, গম) তুলে এনে, গাড়িতে বা অন্য কোনভাবে বহন করে মধ্যবর্তী কোন স্থানে এনে পরিস্কার করে বিভিন্ন শস্য আলাদা করে হোলসেল মার্কেটে বা স্টোরেজে পাঠানো। সেখান থেকে গ্রাহক পর্যায়ে পৌঁছানোর জন্য কাঁচা ফসল বিক্রয় কেন্দ্রে পাঠানো হবে। এই ধাপটিই ইটিএল। অর্থাৎ

এক্সট্রাক্ট- জমি থেকে নানা শস্য তুলে আনা (সব ধরনের সোর্স থেকে ডেটা সংগ্রহ করা) ট্রান্সফরম □ শস্যগুলোকে বাছাই করা, আলাদা আলাদা করে শস্যের ধরন অনুযায়ী বস্তায় ভরা (ডেটা ক্লিনিং, মার্জিং) লোড □ গ্রাহকের কাছে পৌঁছানোর জন্য বাজারে পাঠানোর উপযোগি করা (ডেটা ওয়্যারহাউজিং) কোন ব্যক্তি যেমন নিজে বীজ বুনে জমিতে লাগিয়ে, চাষ করে, ফসল ফলিয়ে সেই ফসল পরিচর্যা করে তারপর টেবিলে বসে খাওয়ার জন্য যা যা করা দরকার তার সবগুলো নিজে করে না, বরং প্রত্যেকটি ধাপে দক্ষতা অনুযায়ী ব্যক্তিবর্গ তাদের অংশটুকু সম্পন্ন করে বলেই ভোক্তাদের টেবিলে খাবার পৌঁছে যায়।

## ২.৯ পরিসংখ্যানে দক্ষতা অর্জন যে কারণে প্রয়োজন

লাইনের মাঝখানে 3.1415927 ধরে নেই  $X = 100 / 200$

$$\frac{d}{dx} \left( \int_a^x f(u) du \right) = f(x)$$

একটি র্যান্ডম নাম্বার -1.0827932

গড় কাকে বলে?  $\bar{X}$  যদি ড্যান্ডম ভ্যারিয়েবল হয় তাহলে এর গড় বের করতে হয়  $\bar{X} = \frac{\sum x}{n}$

### ২.৯.১ ডেটা দিয়ে আমরা আসলে কী করি?

আমরা এই যে ডেটা দিয়ে এত কিছু করার চিন্তা ভাবনা করছি এবং অনেকক্ষেত্রে করেও ফেলছি সেটি কীভাবে সম্ভব হচ্ছে সেটি একটু গভীরে গিয়ে আমাদের ভেদে দেখা দরকার। ডেটা দিয়ে আমরা কোনকিছুর বর্তমান অবস্থা যেমন পরিমাপ করছি তেমনি ভবিষ্যতে কী ঘটবে সেটিও প্রেডিক্ট করছি। এসব কিছুর মূলে রয়েছে ভ্যারিয়েশন বা ভেদ। পরিসংখ্যানের একটু অমূল্য কনসেপ্ট-এর নাম যদি উল্লেখ করতে হয় তাহলে ভ্যারিয়েশনের কনসেপ্টটি একেবারে প্রথমদিকে থাকবে।

আমাদের একটা ধারণা রয়েছে যে পরিসংখ্যান মানেই সারণি আর সংখ্যা। সেটি অবশ্য ভুল নয়। কিন্তু এটি পরিসংখ্যানের অতি ক্ষুদ্র একটি অংশ মাত্র। এটি ঠিক যে পরিসংখ্যানের প্রাথমিক কাজই হলো উপাত্ত থেকে তথ্য বের করে আনা। লক্ষ্য করুন আমি বলেছি উপাত্ত থেকে তথ্য বের করার কথা। আপাতদৃষ্টিতে উপাত্ত আর তথ্য অনেকের কাছে সমার্থক মনে হতে পারে। কিন্তু আমাদের জানতে হবে এদের পার্থক্য কোথায়। ধরা যাক গত বিশ বছরের স্যাটেলাইট ইমেজিং এর মাধ্যমে বাংলাদেশের দক্ষিণাঞ্চলের বনভূমির চিত্র সংগ্রহ করা হয়েছে। এই চিত্রগুলো এবং আনুষংগিক যে ডেটা সংগ্রহ করা হয়েছে সেগুলো শুধুই ডেটা বা উপাত্ত। কিন্তু এই ডেটাকে বিশ্লেষণ করে যখন দেখা যাবে বিশ বছর আগে আমাদের যে পরিমাণ বনভূমি ছিল আজকে তার চেয়ে অনেক কম বনভূমি অবশিষ্ট রয়েছে তখন এটিকে আমরা বলব তথ্য। উপাত্ত থাকলেই সেটি এমনি এমনি কাজের হবে এমনটি নয়। বরং সেটিকে কাজের উপযোগী করা পরিসংখ্যানের অন্যতম কাজ। অবশ্যই উপাত্ত সংগ্রহ এবং তার পরিকল্পনাও পরিসংখ্যানের অংশ।

তাহলে দেখা গেল উপাত্ত মানেই তথ্য নয়। উপাত্তকে বরং খনি থেকে প্রাপ্ত আকরিকের সাথে তুলনা করা যায়। আকরিককে যেমন নানা প্রক্রিয়ার মধ্যে দিয়ে তা থেকে লোহা বের করা হয়। তেমনি উপাত্তকে ধুয়ে মুছে পরিষ্কার করে তা থেকে তথ্য বের করে আনাই পরিসংখ্যানের উদ্দেশ্য। আর সেজন্য যেসব উপকরণ ও পদ্ধতির প্রয়োজন তা আমরা পরিসংখ্যান নামক আপাত: ভীতিকর বিষয় থেকে জানতে পারবো।

উইকিপিডিয়া থেকে- □Statistics is the study of the collection, organization, analysis, interpretation, and presentation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.□

এই সংজ্ঞা থেকে পরিসংখ্যানের একটি দিক সম্পর্কে আমরা ধারণা পাই। সেটি হলো উপাত্তের বিশ্লেষণ, ব্যাখ্যা ও পরিবেশন সম্পর্কিত। পরিসংখ্যানের এই দিকটিকে বলা হয় বর্ণনামূলক পরিসংখ্যান বা ডেসক্রিপটিভ পরিসংখ্যান। এই সংজ্ঞায় পরিসংখ্যানের (আমার মতে) সবচেয়ে গুরুত্বপূর্ণ দিকটিই না বলা রয়ে গেছে। সেটি হলো □ সিদ্ধান্ত গ্রহণে পরিসংখ্যানের ব্যবহার এবং সে সম্পর্কিত দিকটি যাকে ইংরেজীতে আমরা বলি ইনফারেনসিয়াল পরিসংখ্যান। পরিসংখ্যানের যে অংশটি পরিশোধিত উপাত্ত (তথ্য)-কে ব্যবহার করে প্রকল্পের পরীক্ষণ (টেষ্ট অব হাইপোথিসিস) করে এবং তা থেকে সিদ্ধান্ত গ্রহণে (ডিসিসন মেকিং) ভূমিকা রাখে সেই অংশটিকে বলে ইনফারেনসিয়াল পরিসংখ্যান।

## ২.১০ বইটি যাদের জন্য

## ২.১১ বইটি যেভাবে পড়তে হবে

## ২.১২ R পরিচিতি

## ২.১৩ বাংলা কী বোর্ড পরীক্ষা করা হচ্ছে

আমার দেশের নাম বাংলাদেশ। আমি বাংলাদেশে বাস করি

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter ???. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter ???.

Figures and tables with captions will be placed in figure and table environments, respectively.

```
par(mar = c(4, 4, .1, .1))  
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure ২.১. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table ২.১.

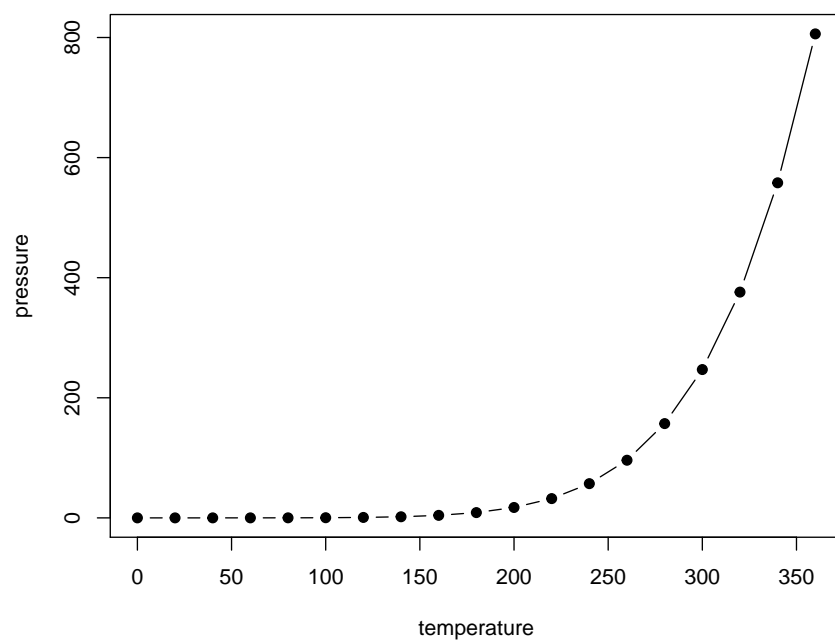
```
knitr::kable(  
  head(iris, 20), caption = 'Here is a nice table বাংলায় সারণির শিরোনাম পরিবর্তন করে দিলাম',  
  booktabs = TRUE  
)
```

You can write citations, too. For example, we are using the bookdown package (Xie, 2018) in this sample book, which was built on top of R Markdown and knitr (Xie, 2015).



সারণী ২.১: Here is a nice table বাংলায় সারণির শিরোনাম পরিবর্তন করে দিলাম

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa



নকশা ২.১: Here is a nice figure বাংলায় ছবির শিরোনাম পরিবর্তন করে দিলাম!

## অধ্যায় ৩

### 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 ক্যারেক্টারিস্টিক 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 মাধ্যমে বিশেষ কোন কাজে লাগানোর উপযোগী করে যা পাওয়া যায় তাকে তথ্য (Information) বলে।

#### ৩.১ ডেটার উদাহরণ

##### ৩.১.১ 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: 73, 38, 86, 23, 41, 61, 33, 41, 39, 50, 47, 37, 76, 20, 41, 69, 51, 48, 58, 55, 47, 64, 57, 45, 21, 38, 27, 1, 38, 46। 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

সারণী ৩.১: ফেইসবুক পোস্টে লাইক-এর সংখ্যা

likes
73
38
86
23
41
61
33
41
39
50

এই ডেটাগুলোকে আপনার কোন একটি কাজের জন্য সরক্ষণ করবেন। তার জন্য আমরা ডেটাকে সাধারণত ট্যাবুলার ফর্মে বা সারণি আকারে সাজাই যেটি দেখতে টেবিল ৩.১-এর মতো হবে।

### ৩.১.২ দৈনিক ইন্টারনেট ব্যবহারের পরিমাণ

আমরা প্রায় সবাই মোবাইল ইন্টারনেট প্যাকেজ ব্যবহার করি। আপনার দৈনিক কতটা ইন্টারনেট ব্যবহার করেন সেটি আপনার একটি বৈশিষ্ট্য যা ইন্টারনেট ব্যবহারের 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 ৩.২-এ দেয়া হল।

## ৩.২ 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। অন্যভাবে বলা যায়

ভ্যারিয়েবল হলো ডেটা ফাইলের মধ্যে কলামকে যে নাম দেয়া হয় সেই নাম।

সারণী ৩.২: দৈনিক ইন্টারনেট ব্যবহারের পরিমাণ

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

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 ভ্যারিয়েবল আছে বলে এরকম ডেটাকে ইউনিভ্যারিয়েট ডেটা (univariate data) বলে। যে ডেটায় দুটি ভ্যারিয়েবল থাকবে তাকে বলবে বাইভ্যারিয়েট (bivariate) ডেটা, এবং দুয়ের অধিক ভ্যারিয়েবল থাকলে সেই ডেটাকে মাল্টিভ্যারিয়েট (multivariate) ডেটা বলে।

এই ডেটাতে তিনটি ভ্যারিয়েবল আছে। ভ্যারিয়েবলগুলো হচ্ছে dates, days, usage। যেহেতু দুয়ের অধিক ভ্যারিয়েবল আছে সেহেতু এই ডেটাকে মাল্টিভ্যারিয়েট ডেটা বলা হবে।

ভ্যারিয়েবল হতে পারে শুধু সংখ্যা, কিংবা তারিখ, কিংবা দিনের নাম (টেবিল ৩.২ দ্রষ্টব্য)। কোন একটি ভ্যারিয়েবল কী রকম মান গ্রহণ করে তার উপর নির্ভর করে ভ্যারিয়েবলের ধরন। মূলত ভ্যারিয়েবল দুই ধরনের: কোয়ান্টিটেটিভ (quantitative) বা সংখ্যাচাচক ভ্যারিয়েবল, 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 ভ্যারিয়েবলের 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 সপ্তাহের 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 বর্ণনামূলক 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 ক্লাস	উচ্চ বিত্ত, মধ্যবিত্ত, নিম্নবিত্ত
-	হালের জনপ্রিয় সাবজেক্ট	স্ট্যাটিসটিক্স, ডেটা সায়েন্স
-	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 উচ্চ কিংবা কোনটি ছোট বা নিচু এরকম করে সাজানো যায় না। তবে অর্থনৈতিক ক্লাস- - 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।

## অধ্যায় ৪

### ডেটা ভিজুয়ালাইজেশন

এই অধ্যায়ে ডেটা ভিজুয়ালাইজেশনের মাধ্যমে আমরা পরিসংখ্যানে মৌলিক কিছু পরভাষার সাথে পরিচিত হব।





## অধ্যায় ৫

### হাইপোথিসিস টেস্টিং

Some significant applications are demonstrated in this chapter.

#### ৫.১ Example one

#### ৫.২ Example two



## অধ্যায় ৬

### সামারি বা সারাংশ

We have finished a nice book.



## গ্ৰন্থসূত্ৰাবলী

Xie, Y. (2015). Dynamic Documents with R and knitr. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). bookdown: Authoring Books and Technical Documents with R Markdown. R package version 0.7.