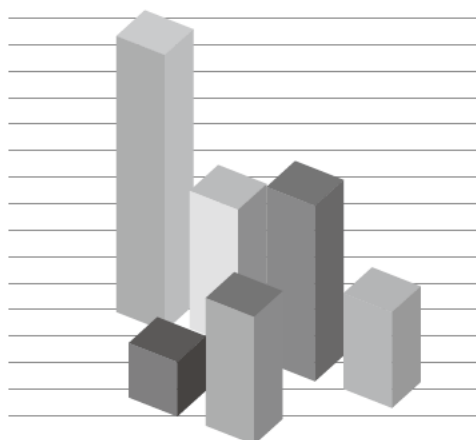


Excel で学ぶデータサイエンス入門講座

No.1

基礎編

監修／株式会社 ROX
執筆／福井 高志



コガク

目次

学習を始めるにあたって.....	1
第1分冊 学習のねらい	2
第1週 ビッグデータとデータの前処理.....	3
1.1 データサイエンティスト / データサイエンスとは?	4
1.1.1 データサイエンティストとは?	4
1.1.2 データサイエンス進展の背景	5
1.2 ビッグデータとは?	7
1.2.1 ビッグデータの定義	7
1.2.2 ビッグデータの特徴	7
1.2.3 ビッグデータの活用	8
1.2.4 データ分析と AI	9
1.3 データの前処理	10
1.3.1 データの前処理とは?	10
1.3.2 Excel におけるデータクレンジング	10
1.3.3 欠損値	12
1.3.4 外れ値	17
1.3.5 表記揺れ	19
1.3.6 データの前処理完了後	22
『まとめと練習問題』.....	24
第2週 基礎統計量	25
2.1 データの種類	26
2.1.1 数量データ	26
2.1.2 カテゴリーデータ	26
2.2 1変量データの概要を掴む	27
2.2.1 度数分布表とは	28
2.2.2 ヒストグラム	31
2.3 基本統計量	34
2.3.1 代表的な値を示す基本統計量	34
2.3.2 代表的な値を示す基本統計量(平均値, 中央値, 最頻値)の関係	36
2.3.3 データの種類と代表的な値を示す基本統計量(代表値)	36
2.3.4 その他の基本統計量(分散と標準偏差, 最大値と最小値)	38
2.3.5 Excel を用いた基本統計量算出の操作	41
『まとめと練習問題』.....	46

第3週 クロス集計	47
3.1 2変数のデータの概要を掴む	48
3.1.1 散布図	48
3.1.2 相関	49
3.1.3 共分散	52
3.2 単純集計表とクロス集計表	53
3.2.1 単純集計とクロス集計	53
3.2.2 クロス集計を使うメリット	54
3.3 ピボットテーブル	55
3.3.1 集計表作成のステップ	55
3.3.2 集計表作成のExcel操作手順	56
3.3.3 ピボットテーブルの便利機能	61
3.4 ピボットグラフ	66
3.4.1 集計データのグラフ化	66
『まとめと練習問題』	67
第4週 グラフによる可視化	69
4.1 グラフによる可視化	70
4.2 様々なグラフ	70
4.2.1 折れ線グラフ	70
4.2.2 棒グラフ	72
4.2.3 円グラフ	74
4.2.4 散布図	77
4.2.5 バブルチャート	78
4.2.6 レーダーチャート	80
4.3 グラフの見た目を整える	82
4.3.1 グラフタイトル	83
4.3.2 軸	83
4.3.3 軸ラベル	86
4.3.4 データラベル	87
4.3.5 凡例	88
4.3.6 行/列の切り替え	89
4.3.7 データ系列の書式設定	89
4.4 2軸グラフ	90
4.4.1 2軸グラフの作成	90
『まとめと練習問題』	93
STEP UP	95
参考文献	96
練習問題の解答	97
索引	98

第1週

ビッグデータとデータの前処理

【学習のポイント】

今週は、ビッグデータの特性についての学習から始めます。データについて理解を深めることは、今般キーワードとなっている IoT や AI を理解することにも繋がります。アクセス可能なデータが爆発的に増加していること、またその流れを受けてデータの重要性がいかに増しているかを実感してください。

後半は、データの質の話です。データが多ければ多いほどいいという訳ではありません。データ分析の質はデータの質によって大きく左右されます。データ分析の前段階として、データをすぐに分析可能な状態にしておくことは、非常に重要なプロセスです。一般的には軽視されがちなこの前処理というプロセスですが、データ分析の実務においては、時には全作業時間の 8 割を占めることもあります。大切な前処理のプロセスの手順を、しっかりと習得してください。

1.1 データサイエンティスト / データサイエンスとは？

1.1.1 データサイエンティストとは？

データ分析といえば、昔から基本のビジネススキルの一つであり、そのスキルを身につけようと多くの人が取り組んできました。更に、最近になってデータサイエンティストが新しい職業として認知されるとともに、ハーバード・ビジネス・レビュー（2013年2月号）では「データサイエンティストほど素敵な仕事はない」と取り上げられるなど、高い注目を集めています。

一般社団法人データサイエンティスト協会では、データサイエンティストを「データサイエンティストとは、データサイエンス力、データエンジニアリング力をベースにデータから価値を創出し、ビジネス課題に答えを出すプロフェッショナル」と定義しており、データサイエンティストに求められるスキルセットとして以下の3つの力を挙げています。

1. ビジネス力 (business problem solving)
課題背景を理解した上で、ビジネス課題を整理し、解決する力
2. データサイエンス力 (data science)
情報処理、人工知能、統計学などの情報科学系の知恵を理解し、使う力
3. データエンジニアリング力 (data engineering)
データサイエンスを意味のある形に使えるようにし、実装、運用できるようにする力

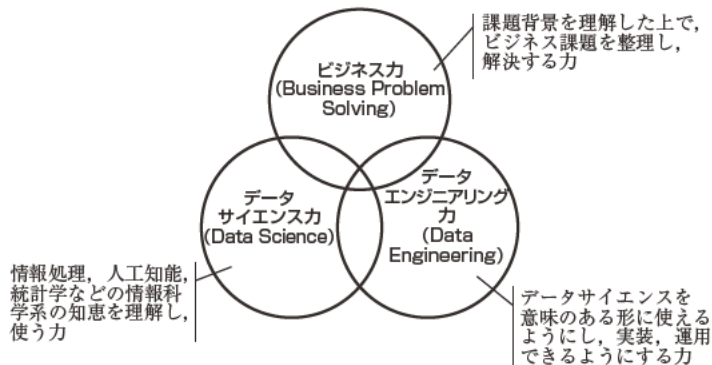


図 1.1 データサイエンティストに求められる3つの力

(出典：一般社団法人データサイエンティスト協会プレスリリース 2014年12月10日)
「データサイエンティスト協会、データサイエンティストのミッション、スキルセット、定義、スキルレベルを発表」図1：データサイエンティストに求められるスキルセット)

本講座は、データサイエンス力としての統計学，データエンジニアリング力として Microsoft Excel をツールとしたデータ分析力を身につけることを目標とした入門講座です。

1.1.2 データサイエンス進展の背景

上記で紹介したデータサイエンティストに求められる力は、従来のデータ分析において求められていた力と比べても、特段の目新しさはありません。では、なぜ今般、データサイエンティストやデータサイエンスなどの言葉がキーワード化しているのでしょうか？その背景の1つとして、データの取得が容易となり、分析対象となるデータの量が爆発的に増えていることが挙げられます。データの取得が容易になった要因は色々ありますが、ここではインターネットの浸透と IoT の進展、この2つを取り上げます。

まず第1の要因はインターネットの浸透です。インターネット環境が普及したことで、私たちがアクセスできる情報量は爆発的に増加しました。今日では、ウェブサイトに関する知見を持たない人でも、ブログなどを通じて情報を発信することができます。やり取りされる情報も、テキストから始まり音声、画像、映像と、データ量は段々と大きくなっており、YouTube ユーザーのアップロードする動画の時間の総計は、2015年時点で1分あたり300時間とも言われています。

第2の要因はIoTの進展です。IoTとは、Internet of Things（モノのインターネット）のことで、あらゆるモノがインターネットに繋がるという「概念」を指します。従来、インターネットに繋がるものと言えばパソコンや携帯電話などで、その利用は画面を介したネットサーフィンやメールなどがほとんどでした。それが、今ではメガネや靴などの身近な製品から、工場内の機械、さらにはスマートシティと呼ばれるインフラレベルまでインターネットに繋がるものが出てきており、インターネットに接続される機器の数は急速に増加を続けています。なお、図1.2の人型下部の数字は一人当たりのIoT機器個数です。

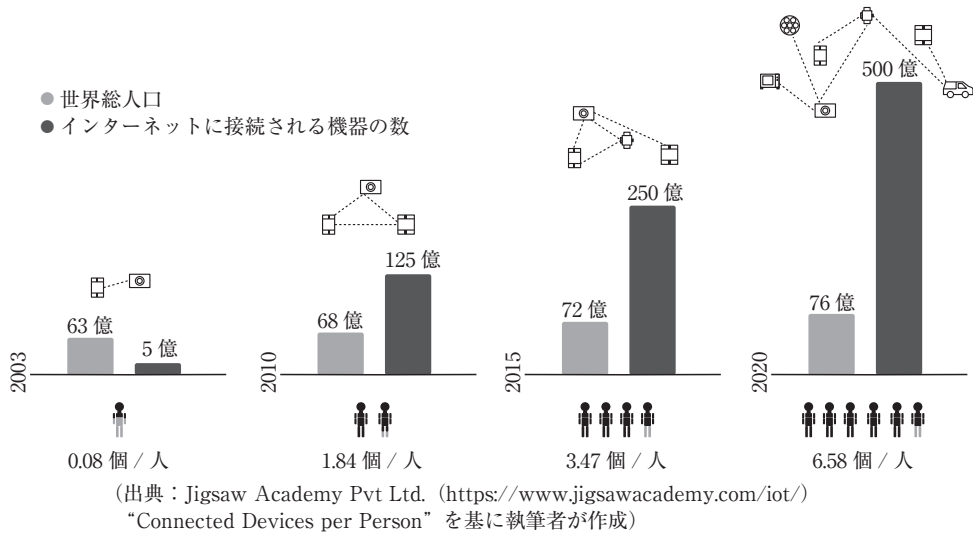


図 1.2 IoT 機器の数

また、データの取得が容易になったことで、取得されるデータの量だけでなく種類も大きく増えました。製品の利用状況や機械の稼働状況、購買行動からインターネットの検索履歴まで、ありとあらゆるデータが日々蓄積されています。

さらには、データの蓄積コストが下がったことも見逃せません。ただデータを貯めておくだけでも、そこにかかるコストは無視できないものですが、クラウドコンピューティングの登場により、そのようなコストも大きく低下しました。

このようにして、これまでは考えられなかった多種多量のデータが取得され、蓄積されるようになりました。これらのデータは、ビッグデータと呼ばれることがあります。

1.2 ビッグデータとは？

1.2.1 ビッグデータの定義

ビッグデータとは、その名の通り巨大なデータ群のことですが、「典型的なデータベースソフトウェアが把握し、蓄積し、運用し、分析できる能力を超えたサイズ」などと定義されることが一般的です。どの程度の量のデータをビッグデータというのか、と言った明確な定義があるわけではありませんが、多くの場合、量的には数十テラバイトから数ペタバイトとされます。

一方で、巨大だけでなく、データが複雑であることも、ビッグデータの特徴としてよく挙げられます。この文脈でよく使われるのが、データを構造化データと非構造化データに大別する分類です。構造化データは、一言で言えば整理がしやすいデータです。Excel や CSV など、行と列の二次元に並んでいるデータがその代表です。逆に非構造化データとは、整理の難しいデータで、規則的な区切りがなく二次元の表形式では表しにくいデータです。非構造化データの代表的な例は、画像や動画、文章などのデータです。この中間で、データ内に規則的な区切りはあるが二次元化しにくい半構造化データもあり、json ファイルや HTML タグ形式が該当します。

従来、データと言えば扱いやすい構造化データがほとんどでしたが、先に述べたようなデータ取得の容易化、取得経路の多岐化によって取得されるようになったデータには非構造化データも多く、これらを総称してビッグデータと呼ぶ傾向があります。

1.2.2 ビッグデータの特徴

ビッグデータの質的な定義、つまりビッグデータが持ちうる特性としては、ガートナーが提唱した次の 3 つの V (Variety, Volume, Velocity) が有名です。

(1) Variety

Variety は、データの多様性です。上で見た通り、データはテキストや音声、画像など様々な情報かつ様々なファイル形式です。