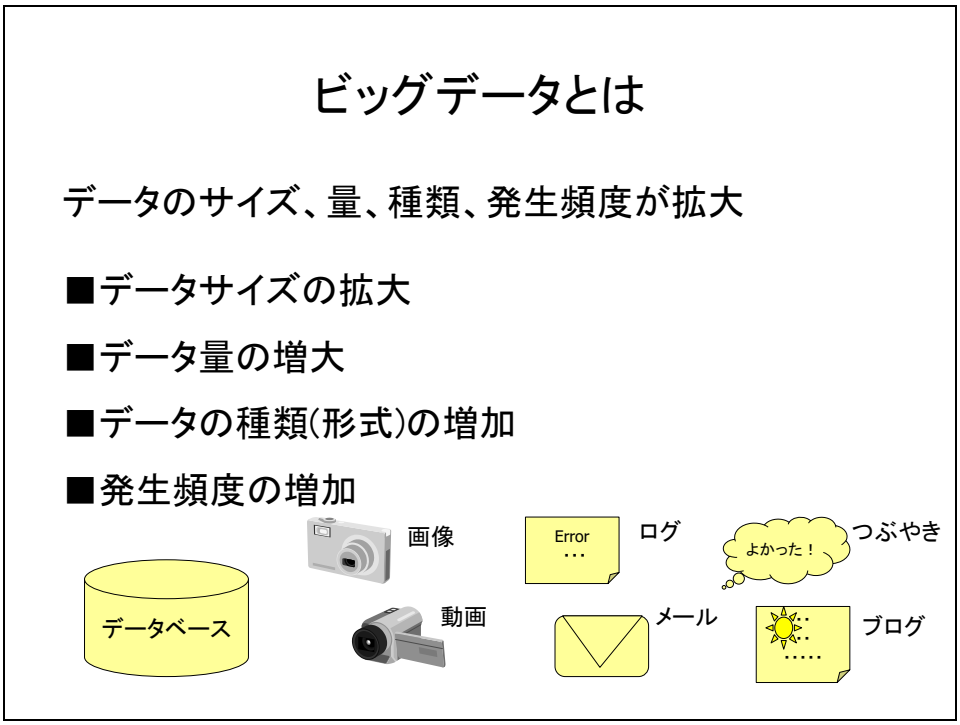
ビジネスデータ分析実践

輸出する際の注意事項

本テキストを非居住者あるいは日本国籍以外の者に転売、貸出または贈与する場合、あるいは外国において非居住者あるいは日本国籍以外の者に転売、貸出または贈与することを目的として日本国外から持ち出す場合は、日本国及び外国の法令等により問題がないか確認を行う必要があり、場合によっては行政庁への許可申請が必要になりますのでご注意ください。

第1章　データ分析の概要

1．1 ビッグデータとは



ビッグデータとは、従来扱ってきたデータよりも、**サイズ**、**量**、**種類(形式)**、**発生頻度**が大きいデータを指します。

■データサイズの拡大

文字、数字、日付などのテキストデータ以外にも、映像、動画などのサイズが大きいデータが増えてきて、1データあたりのサイズが増加しています。

■データ量の増大

業務や生活で必要となるデータに加え、Twitter(つぶやき)、ブログ、ログなどの付加情報が増え、世の中に流通するデータ数は、年々急増しています。

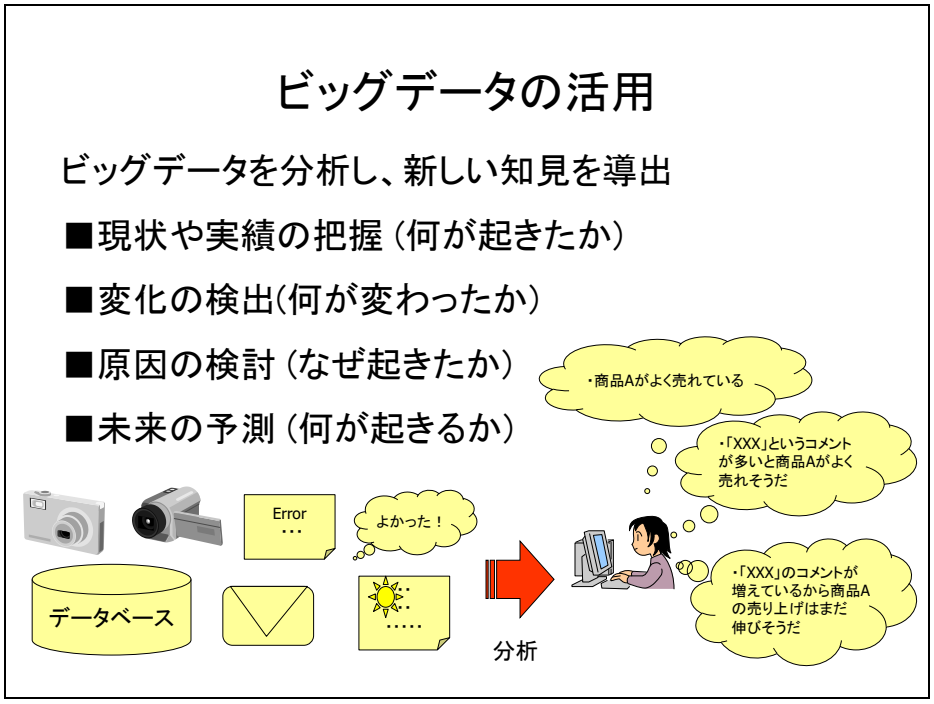
■データの種類(形式)のデータ

文字、数字、日付などのテキストデータ以外にも、文章、映像、動画、ログなどの様々な形式のデータが増えてきています。

■発生頻度の増加

スマートフォンなどのデータ端末の普及率の向上、センサー、防犯カメラなど、データを生成するデバイスの増加にともない、データの発生頻度は年々急増しています。

1．2 ビッグデータの活用



ビッグデータそのものは、単にデータでしかありませんので、ビッグデータをいかに活用していくかということがポイントとなります。ビッグデータの活用方法はいろいろ考えられますが、ビッグデータを分析して新しい知見を導き出し、それを活用することが注目されています。ビッグデータを分析して導出する新しい知見として、以下が考えられます。

・現状や実績の把握
・変化の検出
・原因の仮説
・未来の予測

■現状や実績の把握

ビッグデータを集計し、**過去または現在において何が起きたか**を把握します。今まで知ることができなかったことを把握できるようになれば、それは新しい知見となります。

現状や実績の把握では、特に難しい処理は必要ありません。必要なデータを収集できれば、現状や実績が見られるようになるため、比較的実現は容易になります。

【例】

◆現在の各地域の人口分布、交通状況
◆最近 5 分以内の株取引内容
◆現在の電気使用状況

■変化の検出

ビッグデータを分析することにより、**どのような変化が生じているのか**を把握できるようになります。今まで気づくことができなかった変化を把握できるようになれば、それは新しい知見となります。

変化の検出は、データを収集した後、データ分析により変化が生じたかを分析する必要があります。大きな変化であればすぐ気付くことができますが、小さな変化はデータ分析をしないと、なかなか気づくことができません。

【例】

◆つぶやき数の上昇、下降傾向を検出

◆Web サイトのアクセス速度の変化を検出

◆脈拍、血圧などの変化を検出

■原因の検討

収集したデータを分析することにより、**今まで起きた現象の原因を検討**することができます。今までは気付かなかった**原因と結果の関係**を見つけることができれば、それは新しい知見となります。

原因と結果の関係を検討するには、**データ分析が必要**です。原因と結果の関係は、**大量**のデータを利用するほど精度が上がります。また、**様々な種類**のデータを集めるほど、新しい原因と結果の組み合わせを導出できる可能性が増えていきます。このような特徴もビッグデータ活用が注目されている理由の 1 つです。

ただし、適切なデータ分析をおこなえば、導出した知見の精度は上がりますが、**100%正確な知見を導出することはできない**ことも留意する必要があります。

【例】

◆天気と風邪に関するつぶやき数の関係

◆あるキーワードのつぶやき数と株価との関係

◆天気と使用電力量との関係

■未来の予測

実績値をもとにして、**今後どのような現象が起こるかを予測**することができます。今までできなかった予測ができるようになれば、それは新しい知見となります。未来の予測をするためには、より高度なデータ分析が必要です。そのため、データ分析のエキスパートや高度な分析ツールを利用することが求められ、実現はより難しくなります。また、未来予測の結果は、原因の検討以上に、**精度を上げることが難しい**ことも留意する必要があります。

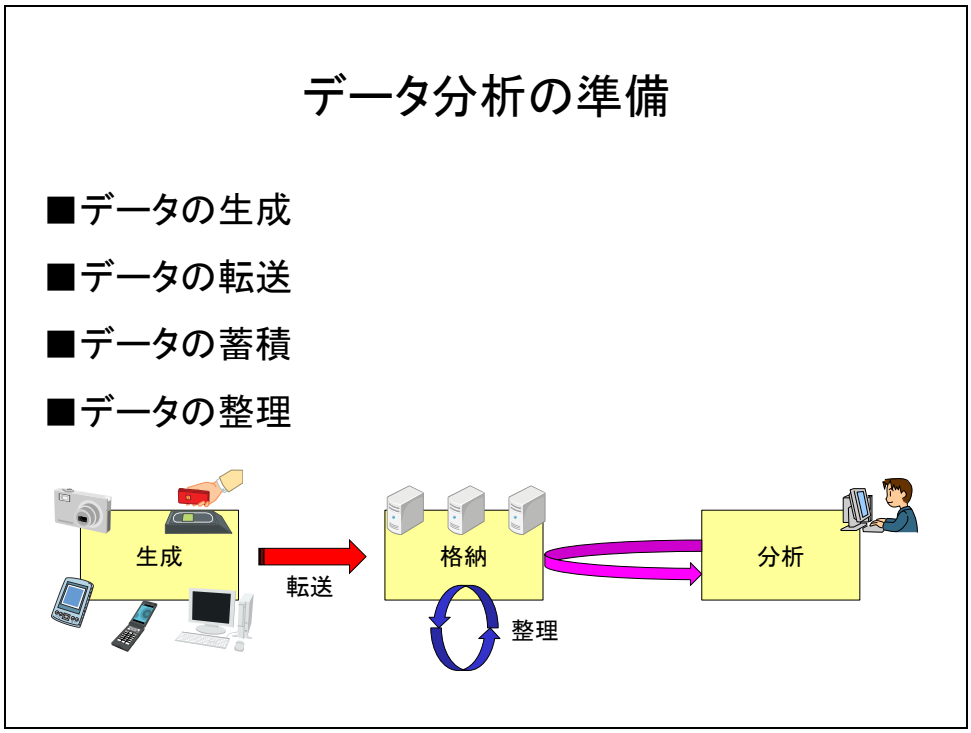
【例】

◆センサーデータと画像、コメントから天気を予測

◆つぶやき、取引状況、経済状況から株価を予測

◆天気、気温などから使用電力量を予測

1．3 データ分析の準備



データ分析を行うためには、分析に必要なデータの準備が必要です。データを準備するためには、「データの生成」「データの転送」「データの蓄積」「データの整理」の具体的な方法を検討します。

■データの生成

データをどのように作成するかを検討します。既存システムのデータ、ログ、ブログなどの既に作成されているデータを利用する場合は、次の「データの転送」を検討します。新規データを利用する場合は、そのデータを生成する手段を考えます。新規データを作成する手段として、おもに以下の方法が考えられます。

◆人手による入力
◆プログラムによる生成
◆センサーなどのマシンより入手

■データの転送

生成されたデータを、分析する環境へ転送する手段を検討します。有線ネットワークを利用する場合は、Ethernet やインターネットなど高速な回線を利用できます。無線通信により、データ転送する場合は、転送速度や通信がつながる場所などの制約がかかる可能性があるため、問題がないかどうかを調査しておく必要があります。

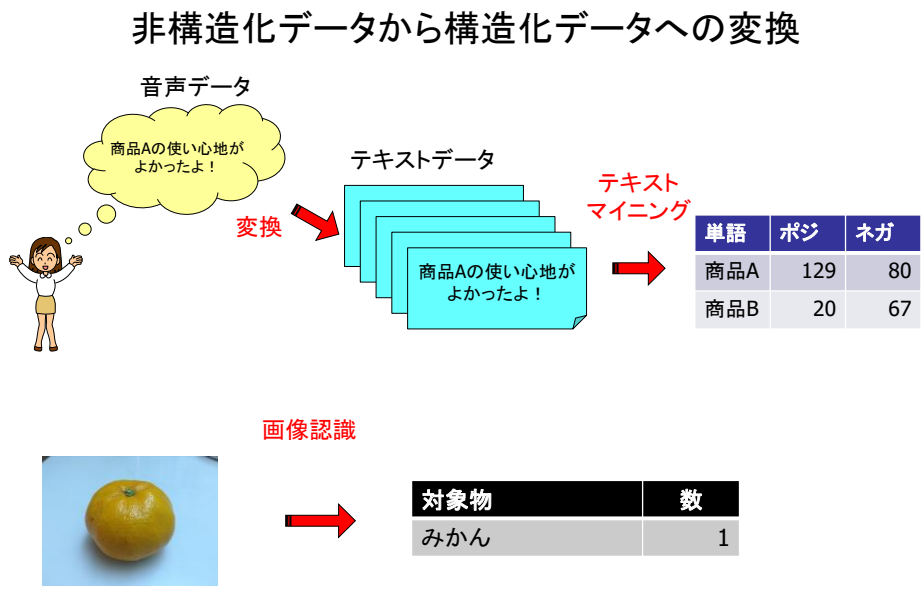
■データの蓄積

生成、転送されたデータを蓄積します。ビッグデータの活用では、データが大きく、多く、非構造、高頻度であるため、それに対応できる方法で蓄積します。ビッグデータの格納先としては、**Hadoop** や **NoSQL** などがおもに利用されています。

■データの整理

生成と転送をしただけでは、分析で利用できないデータがあります。音声、画像、文章などのデータは、そのままでは、統計解析手法によるデータ分析はできません。そのため、音声、画像、文章をデータ分析で利用する場合は、数字や文字データに変換する必要があります。以下に、音声、画像、文章を変換する例を紹介します。

非構造化データ	変換処理の例
音声	音声データをテキストデータに変換し、その後、文章データとして数字や文字に変換。
画像	画像に映っている内容を把握し、その名称、数などのデータに変換。
文章	文章に含まれている単語、単語の出現数、ポジティブ/ネガティブ判定により、文字や数字に変換。



また、不正データや表現の統一などをおこなうためにデータクレンジングを実施します。例えば、「NEC」、「日本電気」、「日本電気(株)」などを「NEC」にまとめます。データクレンジングが終わりましたら、分析手法に適した形式にデータを加工します。

1．4 データ分析の実施、対処

データ分析の実施、対処

データを分析し、価値ある知見を導出。その後、必要に応じて対処を行う

■データ分析ツール

・データ検索・レポートツール / BIツール（OLAPツール）

・データマイニングツール

・人工知能（AI）を利用した分析ツール

分析に必要なデータを準備した後に、データを分析して**知見を導出する手段**を検討します。データ分析ツールとして使用できるものには大きく分けて下記の種類のものが存在します。

■データ検索・レポートツール / BI ツール（OLAP ツール）

データ検索・レポートツールは、収集した大量のデータから、現状や過去に関する**データを見やすく表示（見える化）**するツールです。例えば、商品の売上ランキング、つぶやきランキング、アクセスランキングなどが相当します。また、データ検索やレポート作成に加え、**多次元分析がおこなえる**ツールを BI ツールと呼びます。多次元分析をおこなうことにより、現状を把握するだけではなく、原因の分析もおこなうことができます。

■データマイニングツール

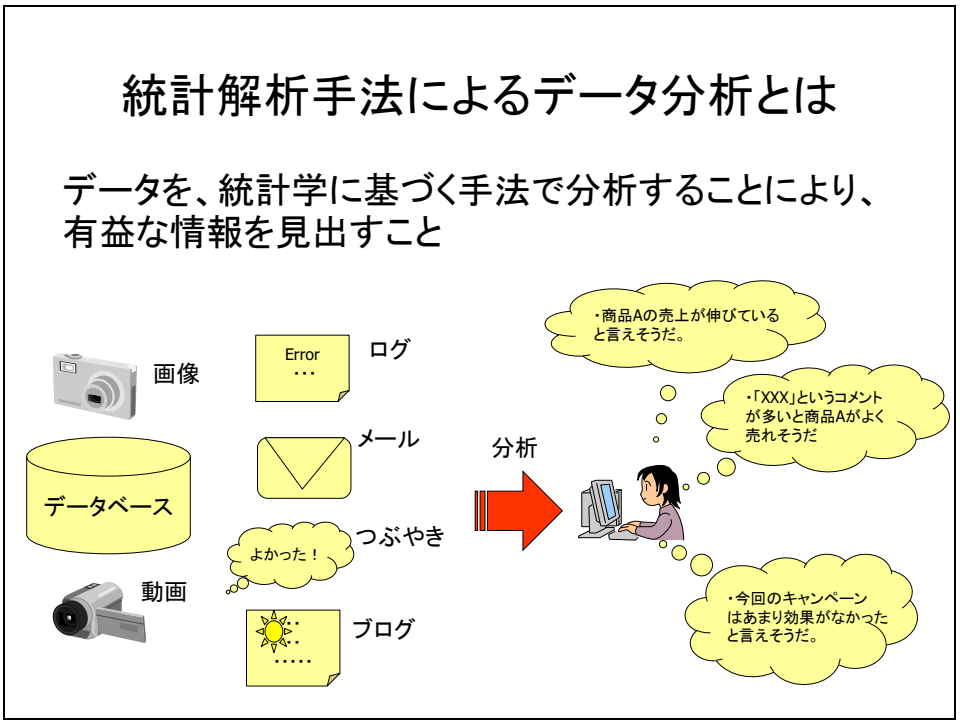
相関分析、判別分析、アソシエーション分析など、**高度な分析がおこなえる**ツールです。高度な分析をすることにより、原因の分析や未来の予測など、より価値のある知見を導出できます。

■人工知能（AI）を利用した分析ツール

近年はデータ分析用のツールとして、AI を用いた分析ツールが利用されるケースもあります。現段階では「データを入れれば何でもやってくれる」ようなものではありませんが、用途、目的を限定することにより、優れた分析結果を提供してくれるものも増えてきています。

分析を行った後は、当初立てた目的を達成できるよう、分析結果を使用して必要な対処（データ可視化の結果を意思決定に反映する、プラント内のアクチュエーター制御のパラメータを変更するなど）を行います。対処後も定期的に評価を行い、必要に応じて新たな分析業務の検討を行います。

1．5 統計解析手法によるデータ分析とは



統計解析手法によるデータ分析とは、統計学に基づく手法でデータ分析することにより、新しい知見を導きだすことです。

データ分析には、統計解析手法を使わない分析もあります。数字やグラフなどを見ることによるデータの特徴を調べる方法です。統計解析手法を用いない分析では、**人間の主観**が入ってしまうこと、**判断ミス**の可能性があることが問題といわれています。

統計解析手法を用いた分析では、**人間の主観が入らない**こと、プログラムなどによる**自動分析**ができること、**見ただけでは分らない知見**が導出できることが利点です。しかし、統計解析手法は今までの実績をもとに分析するため、例外的な事象や新しい事象には不向きと言えます。そのため、データ分析では、人間の知識・経験・カンと統計解析手法による分析を適切に融合することが求められます。

1．6 データ分析手法の検討

データ分析手法の検討

■現状や実績の把握

■変化の検出

■データ分類

■原因分析

■判別

■予測

データ分析をおこなうためには、まず、利用するデータ分析手法を検討します。検討するためには、分析手法の種類とそれぞれの特徴を理解しておく必要があります。おもな分析手法の種類を次頁にまとめます。

種類	分析手法	特徴
現状や実績の把握 (代表値)	平均値、最頻値、中央値、合計、件数、分散、標準偏差、歪度、尖度	データの特徴を１つの代表的な値で表現する。 データの特徴を簡単に表現できるのが利点であるが、細かい特徴を表現できないのが欠点である。
現状や実績の把握 (定型レポート)	ランキング	様々な属性でデータを並べ替え、順位を表示する。
	一般グラフ	棒グラフ、折れ線グラフ、円グラフ、散布図など一般的なグラフでデータの特徴を表現する。
	Ｚチャート	値、累積値、移動合計を表示し、短期的、長期的なデータの傾向を調べる。
	パレート図	値の累積値により、データの構成比を把握する。
	レーダーチャート	各属性値の大きさを中心からの距離で表現する。
現状や実績の把握 (多次元分析)	ドリリング	分析軸の階層情報を利用して、データの特徴を調査
	スライス＆ダイス	集計項目および表示条件を切り換えて、データの特徴を調査
変化の検出	分散分析	データ集合間の違いが、誤差によるものか、必然によるものかを判定
	検定	データ集合に変化が生じているかを判定
データ分類	クラスター分析	似ているデータを同じグループにまとめる分析
	クラス分類	入力されたデータが所属するカテゴリを予測
原因分析	相関分析	2つの項目間の関係の強さを調べる
	アソシエーション分析	複数のデータの組み合わせから、組み合わせの相関ルールや頻出頻度の高い組み合わせを抽出
判別	判別分析	いくつかのグループに分かれたデータをもとに、データが分けられている基準（ルール）を導き出し、未分類のデータがどのグループに所属するかを予測
予測	推定	抽出したデータから、全体の平均値、分散を推測
	回帰分析	実績値をもとに予測式を作成し、値を予測
	時系列分析	時系列データを分析して、その傾向や循環性、季節変動、不規則変動などの特徴をとらえ、将来を予測

下線がある分析手法が、本講義で紹介する分析手法です。

第2章　統計用ソフトウェアＲの基本的な使い方

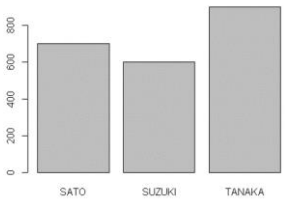
２．１　Ｒ言語概要
２．２　Ｒ基本操作

２．１ Ｒ言語概要

R 言語概要

■統計解析向けのプログラミング言語およびその実行環境
■GNU により開発が進められているオープンソースソフトウェア
■データ分析機能に加え高度なグラフ機能を実装
■豊富な関数やパッケージが存在しており様々な分析に対応

```
> income<-c(700,600,900)
> names(income)<-c("SATO","SUZUKI","TANAKA")
> mean(income)
[1] 733.3333
> barplot(income)
```



R 言語（「R」と呼ばれることもあります）は GNU より開発が進められているオープンソースソフトウェア（OSS）で、統計解析向けのプログラミング言語およびその実行環境です。データマイニングをおこなうツールは高価なものが多いですが、R 言語は OSS 製品であり無償で利用することができます。

R 言語では統計解析やデータマイニング用の多くのパッケージや関数が提供されています。そのため、幅広い分析に対応しており、高度なデータ分析をおこなうことができます。また、グラフ機能も用意されているため、データの視覚化にも優れています。

また、他の統計ソフトウェアとの連携、Java などの他言語からの呼び出し、ODBC によるデータベースアクセスなど拡張性にも優れています。

ただし、R 言語の問題点としては、基本的にメモリ内で処理をおこない一時ファイル処理などは不得意としています。さらにプログラム中でのデータの受け渡しは値渡しのため、大量のメモリを消費することがあります。

しかし、R 言語関連のプロジェクトの中には、これらの問題を解決するための活動をしているものもあり、そこでは大量データ処理用のパッケージ（ff、bigmemory など）も開発されています。

２．２ Ｒ基本操作

ここでは、Ｒ言語の基本操作を確認します。

操作練習１

１．デスクトップのアイコンをダブルクリックして、Ｒを起動します。

２．いくつかの基本的本的なコマンドを実行します。

```
> 10+20
[1] 30

> x <- 10
> y <- 20
> x+y
[1] 30
```

変数に値を代入

変数を使用した計算

３．関数も利用できます。なお、ベクトル(同じデータ型の集合)を定義する際にはベクトルデータの前に「c」を指定します。

```
> z <- 1.546
> round(z,2)
[1] 1.55
> x <- c(1,2,3,4,5)
> sum(x)
[1] 15
> mean(x)
[1] 3
```

z を、小数点第２位に丸めこみます

5 つの数字(1,2,3,4,5)をベクトルとして x に代

x の合計を算出

x の平均を算出

関数の詳細を調べる際は下記のどちらかの方法でヘルプの参照が可能です。

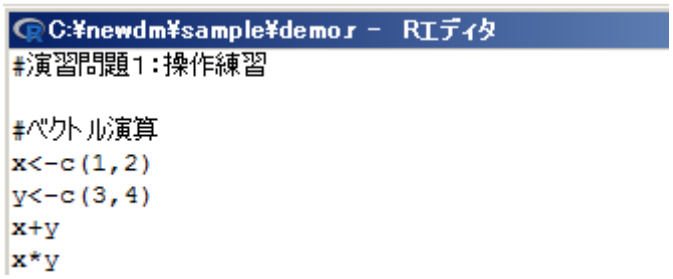
```
> help("round")
> ?round
```

これによって、ブラウザが起動し、該当関数の詳細が表示されます。また、「??検索ワード」と入力することにより、検索ワードを含んだヘルプページの一覧を確認することも可能です。

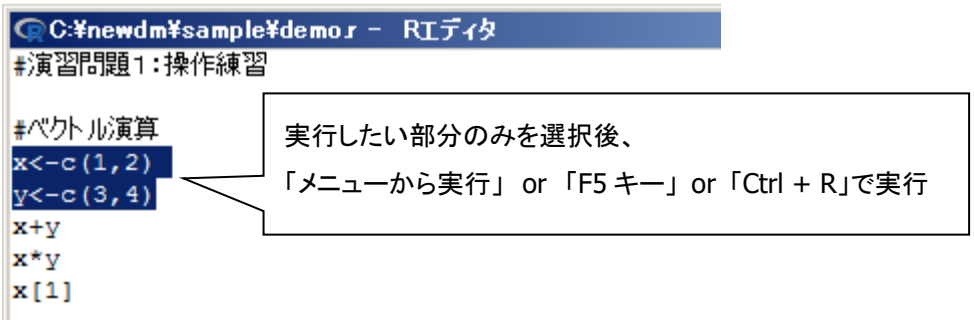
1. 事前に用意されているファイルを利用します。メニュー「ファイル」→「スクリプトを開く」をクリックして、「**c:¥kda¥sample¥demo.r**」を開きます。これによって、スクリプトのエディタ画面が表示されます。

【注意１】テキストと配布スクリプトは Ｃドライブ直下に配布の「kda」フォルダを格納した前提で作成してあります。必要に応じて、テキストの読み替えとスクリプト修正を行ってください。

【注意２】Mac の方でスクリプトファイルを開いた際に文字化けしてしまった場合は、「フォーマット」から「エンコーディングを指定して再度読み込み」で他の文字コードを選択してみてください。



2. スクリプト内より実行したい部分を選択してメニュー「編集」→「カーソル行または選択中の Ｒコードを実行」を選択すると、そのコマンドが実行できます。また、「Ctrl ＋ Ｒ」「F5 キー」でも実行が可能です（Mac の方は Command＋Enter）。



3. この方法を利用して、各コマンドを実行してください。

操作練習２

次の指示に沿って、R の基本的なデータ操作、グラフ化の練習を行いましょう。
下記スクリプトを保存したファイルが「**C:¥kda¥sample¥practice.r**」というファイルで用意してあるので、ファイルを開き、必要な部分を選択して実行しながら内容の確認を行ってください。

１．作業ディレクトリを指定します。

```
> setwd("c:/kda/sample")
> getwd()
```

※R 言語では、ファイルパスに「¥」の代わりに「/」を使用します

setwd 関数：作業フォルダーの指定
getwd 関数：作業フォルダーの確認
スクリプト中「＃」で始まる部分はコメント扱いとなります。

２．作業フォルダーにある CSV 形式のデータファイルを読み込みます。このファイルは、東京、愛知、大阪、福岡の男女別人口統計を集計したものです。
CSV 形式のデータファイルを読み込む際には「read.csv 関数」を使用します。これにより、データを**「表形式のデータフレーム型（data.frame）」**で取り込むことが可能です。

```
> ds <- read.csv("jinkou.csv", header=T)
```

read.csv 関数　第 1 引数：CSV ファイル名
read.csv 関数　header オプション：1 行目が列名の場合「T」

３．読み込んだデータを確認します。

```
> ds
```

読み込んだデータはデータ数も多いため、直接データを見ても内容を確認するのが困難です。
そのため、別の方法でデータを確認します。

４．次のコマンドを用いて、読み込んだデータを調べてください。

先頭部のデータを表示

```
> head(ds)
```

最終部のデータを表示

```
> tail(ds)
```

年の範囲（最小値と最大値）を確認

```
> range(ds$year)
```

データフレームの列の指定方法
変数名$列名

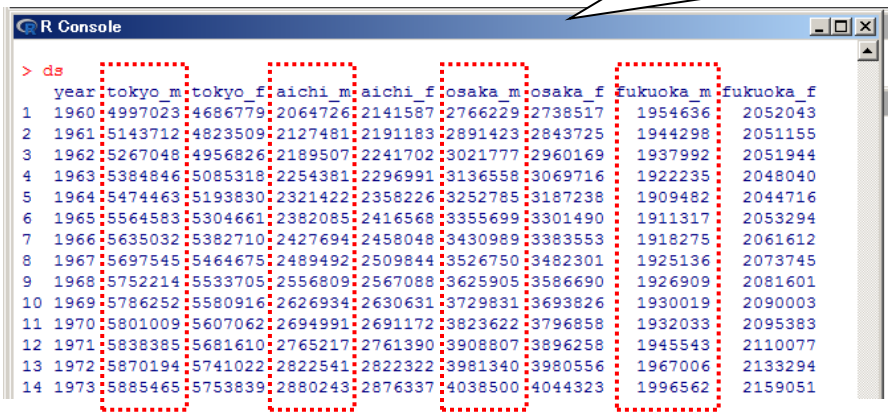
head、tail、range といったコマンドを使用することにより、データの先頭数行分、最終部数行分や、値の範囲を確認することが可能です。データを読み込み、分析を行う前に、「データがどのような内容で、どのような形で存在するのか」を確認するために使用するとよいでしょう。

５. 統計情報の確認をおこないます。

各年度の男性の人口を求め、male 変数に代入します。
その後、male の中身を確認し、mean 関数を使用して、男性の平均人口を調べてください。

```
> male <- ds$tokyo_m+ds$aichi_m+ds$osaka_m+ds$fukuoka_m
> male
> mean(male)
```

点線で囲まれた「男性の人口」部分を取り出し、変数 male に代入しています。




```
> ds
   year tokyo_m tokyo_f aichi_m aichi_f osaka_m osaka_f fukuoka_m fukuoka_f
1  1960 4997023 4686779 2064726 2141587 2766229 2738517   1954636   2052043
2  1961 5143712 4823509 2127481 2191183 2891423 2843725   1944298   2051155
3  1962 5267048 4956826 2189507 2241702 3021777 2960169   1937992   2051944
4  1963 5384846 5085318 2254381 2296991 3136558 3069716   1922235   2048040
5  1964 5474463 5193830 2321422 2358226 3252785 3187238   1909482   2044716
6  1965 5564583 5304661 2382085 2416568 3355699 3301490   1911317   2053294
7  1966 5635032 5382710 2427694 2458048 3430989 3383553   1918275   2061612
8  1967 5697545 5464675 2489492 2509844 3526750 3482301   1925136   2073745
9  1968 5752214 5533705 2556809 2567088 3625905 3586690   1926909   2081601
10 1969 5786252 5580916 2626934 2630631 3729831 3693826   1930019   2090003
11 1970 5801009 5607062 2694991 2691172 3823622 3796858   1932033   2095383
12 1971 5838385 5681610 2765217 2761390 3908807 3896258   1945543   2110077
13 1972 5870194 5741022 2822541 2822322 3981340 3980556   1967006   2133294
14 1973 5885465 5753839 2880243 2876337 4038500 4044323   1996562   2159051
```

同様に女性の人口を求め変数 female に代入し、変数の内容、平均人口を確認してみましょう。

```
> female <- ds$tokyo_f+ds$aichi_f+ds$osaka_f+ds$fukuoka_f
> female
> mean(female)
```

6. 各年度の男女の合計人口の推移をグラフにプロットします。グラフのプロットには plot 関数を用います。

```
> t1 <- male+female
> plot(ds$year,t1,type="l",xlab="年",ylab="人口")
```

plot 関数：データをグラフに表示

第 1 引数：X 軸のデータ

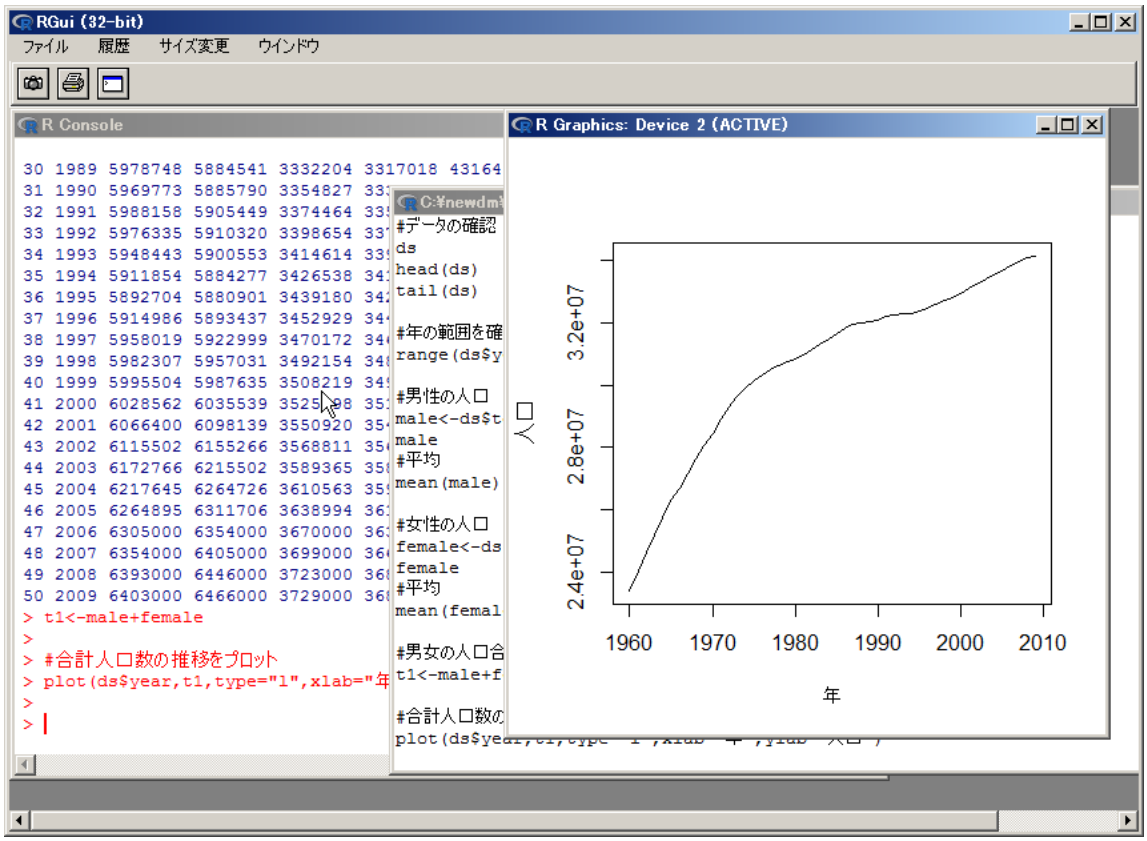
第 2 引数：Y 軸のデータ

type オプション：グラフの種類（線：l、点：p、ヒストグラム：h）

xlab オプション：X 軸の名前

ylab オプション：Y 軸の名前

これによって、画面にグラフが表示されます。



操作練習３

次の指示に沿って、Ｒを使用した基本的なデータ操作を行いましょう。ここでは、Ｒを使用したデータ分析で多く用いられる「データフレーム」に関しての簡単な操作練習を行います。

下記スクリプトを保存したファイルが「**C:¥kda¥sample¥practice2.r**」というファイルで用意してあるので、ファイルを開き、必要な部分を選択して実行しながら内容の確認を行ってください。

■データフレームの定義

データフレームの生成は、次のように行います。「列名 = ベクトル」の列を並べて指定します。各々のベクトルの要素数は同じである必要があります。

```
> (a <- data.frame("製品" = c("テレビ","冷蔵庫","洗濯機"),
                   "単価" = c(20,10,15),
                   "販売数" = c(1000, 400, 350)))
    製品 単価 販売数
1 テレビ   20   1000
2 冷蔵庫   10    400
3 洗濯機   15    350
```

■要素の参照

要素の参照は、添え字や名前を用いて行います。下記にいくつかパターンを掲載します。

```
> a[1, 1]          # 1 行目、1 列目のデータを取得する
[1] テレビ
Levels: テレビ 洗濯機 冷蔵庫
> a[ , 2]          # 全行、2 列目のデータを取得する
[1] 20 10 15
```

■データフレームに新規列を追加する

データフレームに列を追加するには、transform 関数を利用します。

```
> (a <- data.frame("製品" = c("テレビ","冷蔵庫","洗濯機"),
                   "単価" = c(20,10,15),
                   "販売数" = c(1000, 400, 350)))

    製品 単価 販売数
1 テレビ   20   1000
2 冷蔵庫   10    400
3 洗濯機   15    350

> (b <- c("AAA","BBB","CCC"))

[1] "AAA" "BBB" "CCC"

> (c <- transform(a,"追加列"=b))

    製品 単価 販売数 追加列
1 テレビ   20   1000    AAA
2 冷蔵庫   10    400    BBB
3 洗濯機   15    350    CCC
```

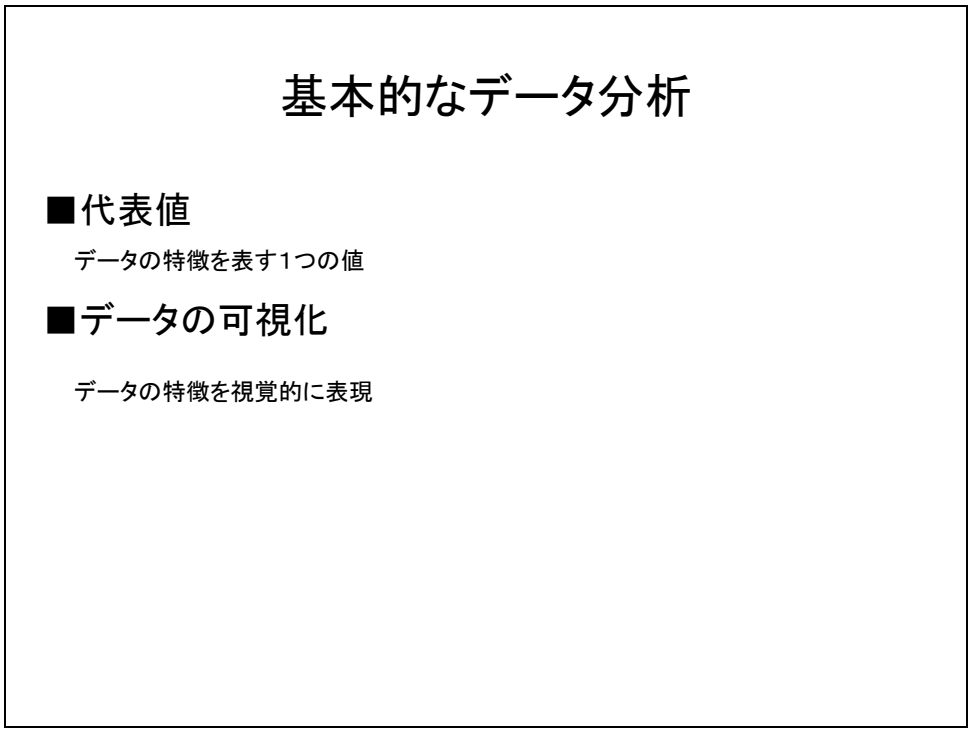
■作業ディレクトリの変更、確認、データのファイル出力

データフレームの中身をファイルに出力するには、write.table 関数を利用します。第 1 引数にデータフレーム名を指定し、その他引数でファイル名（file）、引用符（quote）、セパレータ(sep)、列名有無(col.names)、行名有無(row.names)などを指定可能です。

```
> setwd("c:/kda/sample/")

> getwd()

[1] "c:/kda/sample"

> write.table(c,file="output.txt",quote=F,sep=",",row.names = F,col.names = T)
```

第3章　基本的なデータ分析

3．1 基本的なデータ分析



本章では、データ分析の中では基本となる「代表値」と「定型レポート」について紹介します。この2つは、データの準備ができれば比較的簡単に実施できるデータ分析です。データ分析を始める場合は、まずこの2つから始めて、データの特徴を把握します。

■代表値

「平均値」「標準偏差」「中央値」など、データの特徴を 1 つの代表的な値を算出します。データの特徴を簡単に表現できるのが利点ですが、細かい特徴が分からないのが欠点です。

■データの可視化

棒グラフ、散布図など、データを見やすい形で表現します。視覚的にデータの特徴を理解できるのが利点ですが、解釈には時間がかかること、機械的に判断できないこと、複数のデータを比較しづらいことが欠点です。

3．2 代表値（基本統計量）

代表値（基本統計量）

■平均
■中央値
■最頻値

■最大値、最小値、範囲
■合計
■件数

■分散
■標準偏差

■クォンタイル点

データの特徴を表す代表値には、いくつか種類があります。ここでは、おもな代表値の特徴を紹介します。

■平均値

平均値とは、合計値をデータの個数で割った値です。データが、だいたいどのくらいの値になるかを表現する場合に使用します。日経平均株価、月ごと平均温度など、平均値は、代表値のなかでも非常に馴染みに深い値です。平均値は、**値の大きさがおよそどのくらい**かを把握するために使用します。

■中央値

中央値とは、複数あるデータを値の順に並べて**真ん中の順番**になった値のことです。平均値では、非常に大きい値や小さい値(外れ値)が存在した場合に代表値として不適切になる場合がありますが、中央値は外れ値の影響を受けづらいという特徴があります。

■最頻値

最頻値とは、複数あるデータの中で**一番多く出現する値**のことです。データがどの値になりやすいかという視点では、平均値より適した代表値になります。

■分散

分散とは、**個々の値と平均の差(偏差)を2乗**した値を**合計**し、それを**データ件数で割った値**です。分散により、データが平均値付近に密集しているのか、または平均値から大きく離れて散らばっているのか（データのばらつき具合）を判断することができます。分散が小さいほど値のばらつきが少なく、分散が大きくなると、値のばらつきが大きいといえます。

■標準偏差

標準偏差とは、**分散の値の非負の平方根**をとった値です。分散同様、データのばらつき具合を判断することができます。分散は、個々の値と平均の差を2乗した値の平均であるため、個々の値が平均からどのくらい離れているかを同じスケールで表現できません。そのため、分散の平方根を取った標準偏差により、個々の値と平均値との差を同じスケールで表現します。

■最大値、最小値、範囲

データの最大値、最小値を求めることにより、データの範囲を把握できます。

■合計

代表値の算出に使用したデータの合計値です。データの全体量が把握できます。

■件数

代表値の算出に使用したデータ件数です。データセット間で代表値を比較する場合は、それぞれの件数の差が小さいほうが理想的です。

■クォンタイル点

クォンタイル点とは、対象のデータを 4 分割したときの、最小値、第一四分位（25%地点）、中央値（50%地点）、第三四分位（75%地点）、最大値を示すものです。

R では、関数などを使用して容易に基本統計量を求めることが可能です。ここでは、下記のようなデータを使用して、基本統計量を求めていきます。

【サンプルデータの生成】

```
data <- c(1,1,2,3,4,5,5,5,6,6,6,7,7,7,7,7,8,9,10,10,80)
```

【平均、中央値、最頻値】

```
> mean(data)       # 平均
[1] 9.333333

> median(data)     # 中央値
[1] 6

> table(data)      # table 関数で集計表を作成し、一番多いものが最頻値
data

 1  2  3  4  5  6  7  8  9 10 80
 2  1  1  1  3  3  5  1  1  2  1
```

【分散、標準偏差】

```
> var(data) # 分散（不偏分散：分散の計算時に分子を「データ件数－１」で割ったもの）
[1] 268.7333

> sd(data) # 標準偏差（不偏標準偏差）
[1] 16.39309
```

【最大値、最小値、範囲、合計】

```
> max(data) # 最大値
[1] 80

> min(data) # 最小値
[1] 1

> range(data) # 範囲
[1]  1 80

> sum(data) # 合計
[1] 196
```

【クォンタイル点】

```
# クォンタイル点を求める場合、quantile 関数を使用するか、summary 関数を使用。

> quantile(data) # 四分位数の確認が可能
  0%  25%  50%  75% 100%
   1    5    6    7   80

> summary(data) # 四分位数に加え、平均値も同時に確認可能
 Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
 1.000   5.000    6.000   9.333    7.000    80.000
```

3．3 代表値を見るときのポイント

収集したデータの代表値を算出した後は、代表値からデータがどのような特徴をもっているかを理解し、そのデータに対してどのようなアクションを起こしていくかを検討する必要があります。

まず、代表値を調べる前に、データがどのような特徴であるべきかを考えます。一般的にデータは、「値が高いほど良い」、「値が低いほど良い」、「特定の値であることが望ましい」のいずれかになります。

理想とするデータの特徴	代表例
値が高いほど良い	売上数、売上金額、利益、勝利数、支持率、来訪数、
値が低いほど良い	費用、消費電力、敗戦数、エラー数、バグ数、犯罪数
特定の値が望ましい	アクセス速度、脈拍、血圧

理想とするデータの特徴ごとに代表値の見るべきポイントは、以下のようになります。

理想とするデータの特徴	代表値
値が高いほど良い	平均値、合計値、最大値が**高いほど理想的**である。 **分散、標準偏差が高い**場合は、その値を変化させる要因が存在する可能性があるため、**変化させる原因**を調査する。 分散、標準偏差が低い場合は、その値を上昇させるのが難しいと考えられるため、現状維持を検討する。
値が低いほど良い	平均値、合計値、最大値が**低いほど理想的**である。 **分散、標準偏差が高い場合**は、その値を変化させる要因が何かある可能性があるため、**変化させる原因**を調査する。 分散、標準偏差が低い場合は、その値を下降させるのが難しいと考えられるため、現状維持を検討する。
特定の値が望ましい	**平均値が理想の値**になっているかを確認し、そうでない場合は、理想の値になるような施策を実施する。 理想の値になっている場合は、**分散、標準偏差を確認**し、これらが高い場合は、**値が変化する要因**を調査する。 平均値が理想の値になっており、分散、標準偏差が十分低い場合は、**平均値および分散が変化しないかを監視**する。

3．4 Rによるデータの可視化

Rによるデータの可視化

■棒グラフ
■折れ線グラフ
■ヒストグラム
■散布図
■円グラフ
■箱ひげ図

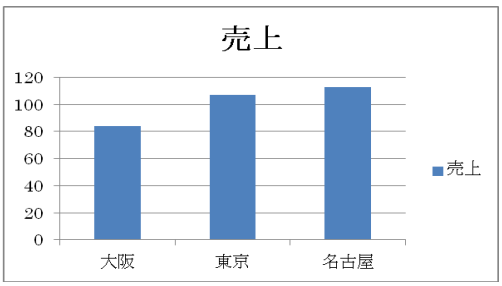
代表値は、値だけでデータの特徴を表現できるので、簡単にデータの特徴を説明する場合には便利ですが、特徴の詳細を把握することができません。特徴の詳細を理解するためには、データの可視化を行います。データの可視化（グラフ化）の種類には、様々なものがあります。本資料では、代表的なおもなグラフとそれぞれの特徴を紹介します。

3．4．1 棒グラフ

【入力データ】管理対象項目と比較する値(評価項目)を準備する。

【出力データ】評価項目の大小関係の度合いを、視覚的に把握できる。

都道府県	売上数
大阪	84
東京	107
名古屋	113



入力データ　　**出力データ**

【分析結果の評価のポイント】

・データの差が、どれくらいあるかを把握する。

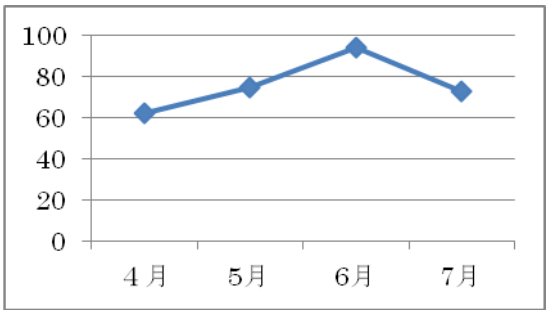
・差がある場合は、差が生まれた原因を推測する。

3．4．2 折れ線グラフ

【入力データ】推移を把握したい項目(評価項目)と推移軸の項目を準備する。

【出力データ】評価項目の値の推移を、視覚的に把握できる。

月	売上数
4 月	62
5 月	75
6 月	94
7 月	73



入力データ　　**出力データ**

【分析結果の評価のポイント】

・推移軸の変化により、評価項目がどのように推移するかを把握する。

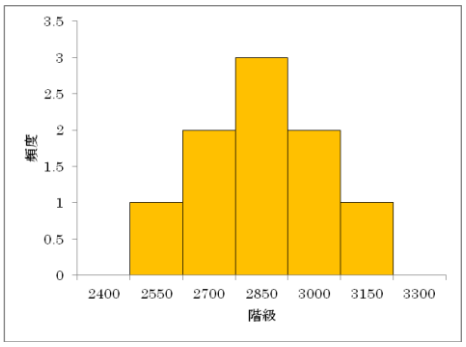
・評価項目の推移の原因を推測する。

3．4．3 ヒストグラム

【入力データ】分布を把握したい項目を準備する。

【出力データ】収集したデータを階級に分け、階級ごとの件数を棒グラフで表現することにより、データの分布状況を視覚的に把握できる。

マンション価格
2550
2700
中略
3000
3150



入力データ　　**出力データ**

【分析結果の評価のポイント】

・管理対象項目の比率が、全体のどれくらいを占めるかを把握する。

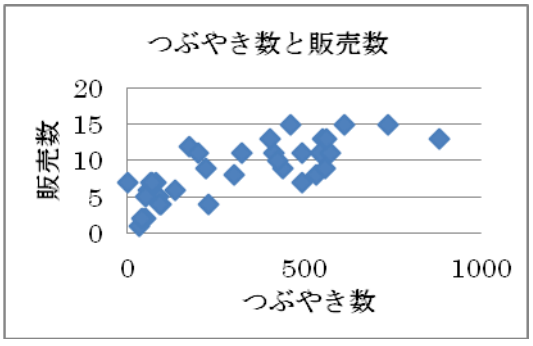
・比率の差の原因を推測する。

3．4．4 散布図

【入力データ】関連があるかを調べたい項目の組み合わせを準備する。

【出力データ】2 つの項目の関連性を感覚的に把握できる。

つぶやき数	販売数
228	4
87	5
492	11
58	6
…	…



入力データ　　**出力データ**

【分析結果の評価のポイント】

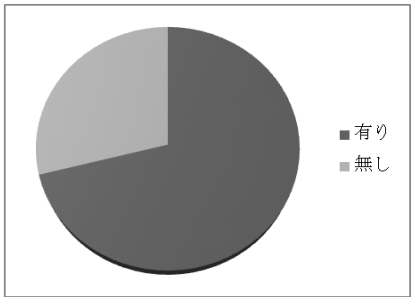
・散布図の分布から、2つの項目間に関連があるかを調べる。

3．4．5 円グラフ

【入力データ】比率を把握したい項目（管理対象項目）と比率を算出する項目を準備する。

【出力データ】管理対象項目の全体比率を感覚的に把握できる。

プールの有無	売上
有り	216
無し	88



入力データ　　**出力データ**

【分析結果の評価のポイント】

・管理対象項目の比率が、全体のどれくらいを占めるかを把握する。

・比率の差の原因を推測する。

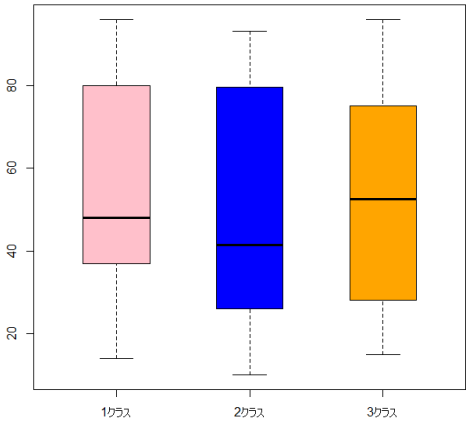
3．4．6 箱ひげ図

【入力データ】四分位数を把握したい項目を準備する。

【出力データ】基本統計量である第 1 四分位から第 3 四分位までの範囲を図示する。

クラス	成績
1	80
1	62
省略	
2	40
2	19
省略	
3	15
3	25
省略	

入力データ



出力データ

【分析結果の評価のポイント】

・個別値が、時系列の値によりどのように推移するかを確認する。

・個別値累計が、直線的に増えているかを確認する。直線から外れている場合は、個別値が増えた(または減少した)原因を推測する。

・移動年計がどのように変化しているかを確認し、長期的な傾向を把握する。

■R を使用した棒グラフの作成

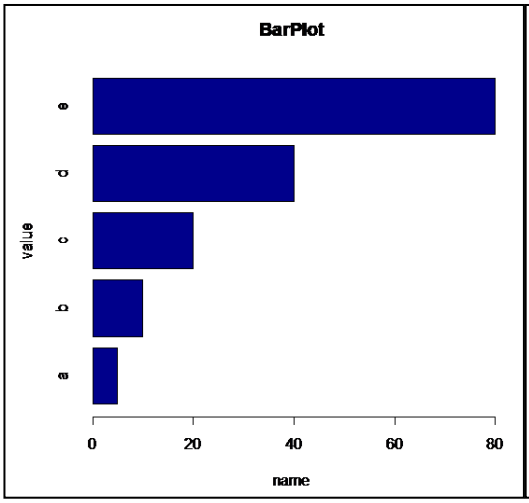
棒グラフを作成するには barplot 関数を使用します。下記の例では、data という変数にベクトルとしてデータを代入後、barlopt 関数を使用して棒グラフを作成しています。

```
> data=c(5,10,20,40,80)
> barplot(data,main="BarPlot",xlab="name",ylab="value",col="darkblue",
 horiz=T, names.arg=c("a","b","c","d","e"))
```

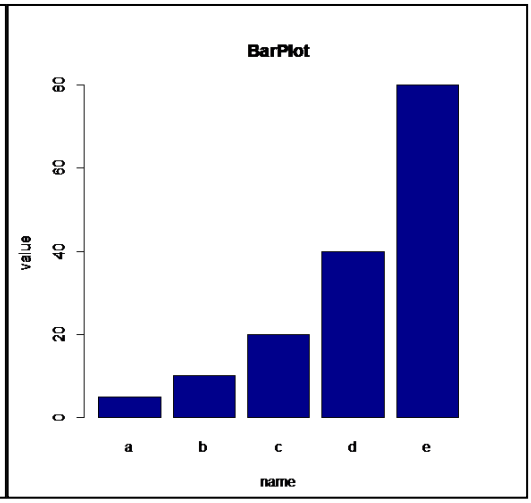
col オプション:棒グラフの色を指定可能

horiz オプション:T にすると横棒グラフ、F にすると縦棒グラフを描画。省略値は F。

names.arg オプション:各項目の名前を設定可能



horiz=T



horiz=F

処理対象のデータが次のような場合を考えます。

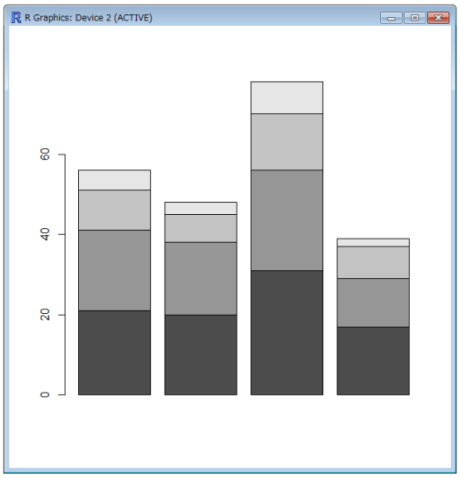
	A 店	B 店	C 店	D 店
10 代以下	21	18	31	17
20 代	20	18	25	12
30 代	10	7	14	8
40 代以上	5	3	8	2

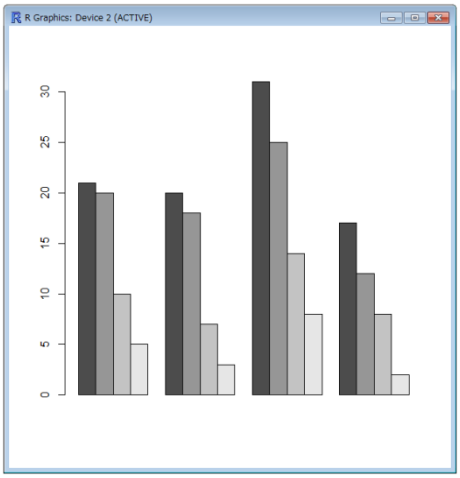
店舗ごとの売上を、年代別に積み上げた棒グラフとして描画するには、次のように行います。

```
> s2 <- matrix(c(21, 20, 10, 5, 20, 18, 7, 3, 31, 25, 14, 8, 17, 12, 8, 2),
ncol=4, nrow=4)
> barplot(s2)
```

年代別の売上を、それぞれ棒グラフとして横に並べるには、次のように行います。

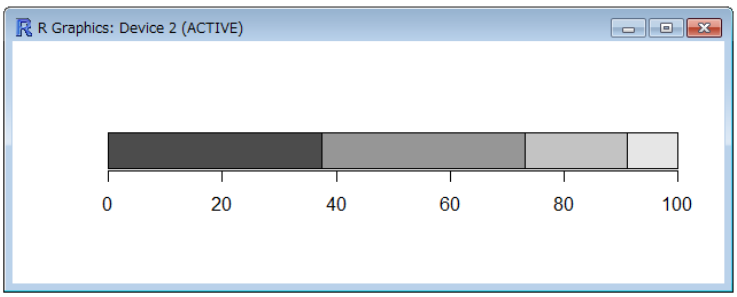
```
> barplot(s2, beside=TRUE)
```





1 列だけの行列を対象にすれば、帯グラフが描画できます。

```
> a <- c(21, 20, 10, 5)      # A 店の売上を対象
> s3 <- as.matrix(a / sum(a) * 100)    # 値を%に換算して行列に変換
> barplot(s3, horiz=TRUE)              # 横向きの棒グラフ(帯グラフ)
```



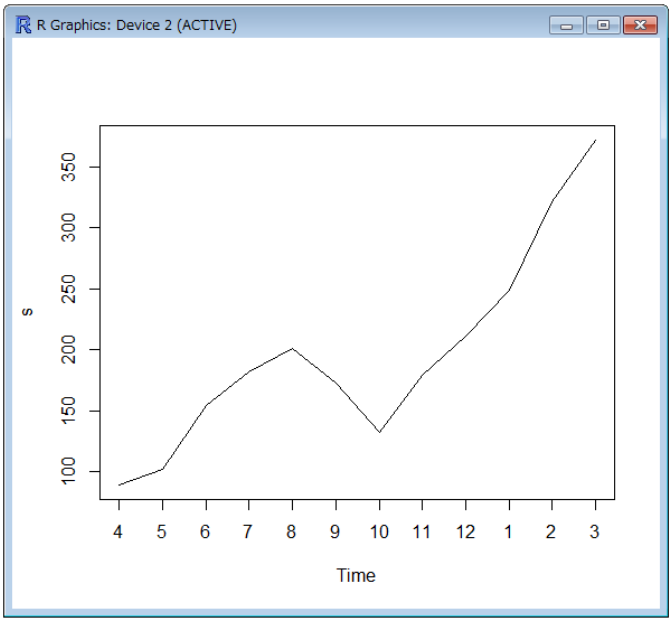
■R を使用した折れ線グラフの作成

折れ線グラフを描画するためには、plot 関数を使用します。下記の例では、「ある商品の売上の推移を示したデータ」の売り上げ状況を折れ線グラフで描画しています。

【ある商品の売り上げの推移を示したデータ】

2014/04	2014/05	2014/06	2014/07	2014/08	2014/09	2014/10	2014/11	2014/12	2015/01	2015/02	2015/03
89	102	154	182	201	173	132	179	211	249	321	372

```
> s <- ts(c(89, 102, 154, 182, 201, 173, 132, 179, 211, 249, 321, 372), start=1, end=12, frequency=1)
> plot(s, xaxt="n", xaxp=c(1, 12, 11))
                          # x 軸の目盛りを表示せず、X 軸の両端の値を 1、12 にする
> axis(side=1, at=1:12, labels=c(4:12, 1:3))
                          # x 軸の 1～12 に、ラベル 4,5,6,...12,1,2,3 を割り当てる
```

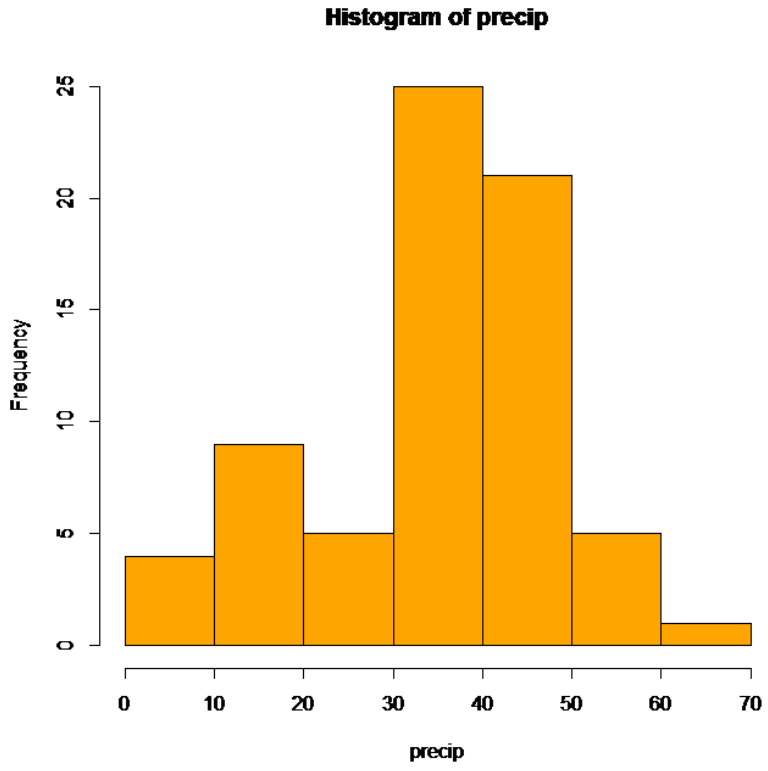


■R を使用したヒストグラムの作成

ヒストグラムを描くには、hist 関数を使用します。引数に、数値の要素からなるベクトルを指定し、この値のある範囲を x 軸、その範囲の要素の個数を y 軸としたヒストグラムを作成します。

ここでは、R をインストールするとデフォルトで利用可能な「アメリカの都市の年間降水量」の関係を示した precip データセットを用いて、降水量の分布状況をヒストグラムとして作成します。

```
> head(precip, n = 3)
 Mobile  Juneau Phoenix
   67.0    54.7     7.0
> hist(precip,col="orange")
```

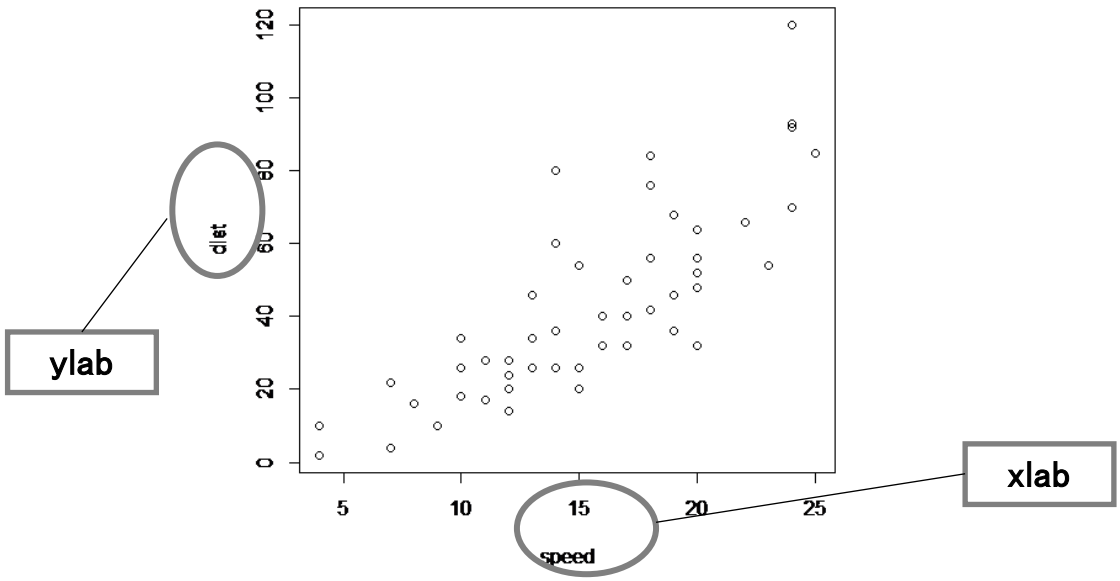


R の標準関数には最頻値を求めるものがないため、最頻値の確認に、ヒストグラムを利用しても良いでしょう。

■R を使用した散布図の作成

散布図を描くには plot 関数を使用します。実行すると、別ウィンドウが開いて、図が表示されます。下記の例では、R をインストールするとデフォルトで使用可能な「車の速度と停止までの距離」の関係を示した cars データセットを用いて、速度と停止距離の関係を散布図として可視化します。xlab 引数に x 軸の名称、ylab 引数に y 軸の名称を指定しています。

```
> head(cars,n=3)  # cars データの初めの 3 件を表示
  speed dist
1     4    2
2     4   10
3     7    4

> x <- cars[,1]   # x 軸に車の速度
> y <- cars[,2]   # y 軸に停止時間
> plot(x, y, xlab = "speed", ylab = "dist") # plot 関数で可視化
```



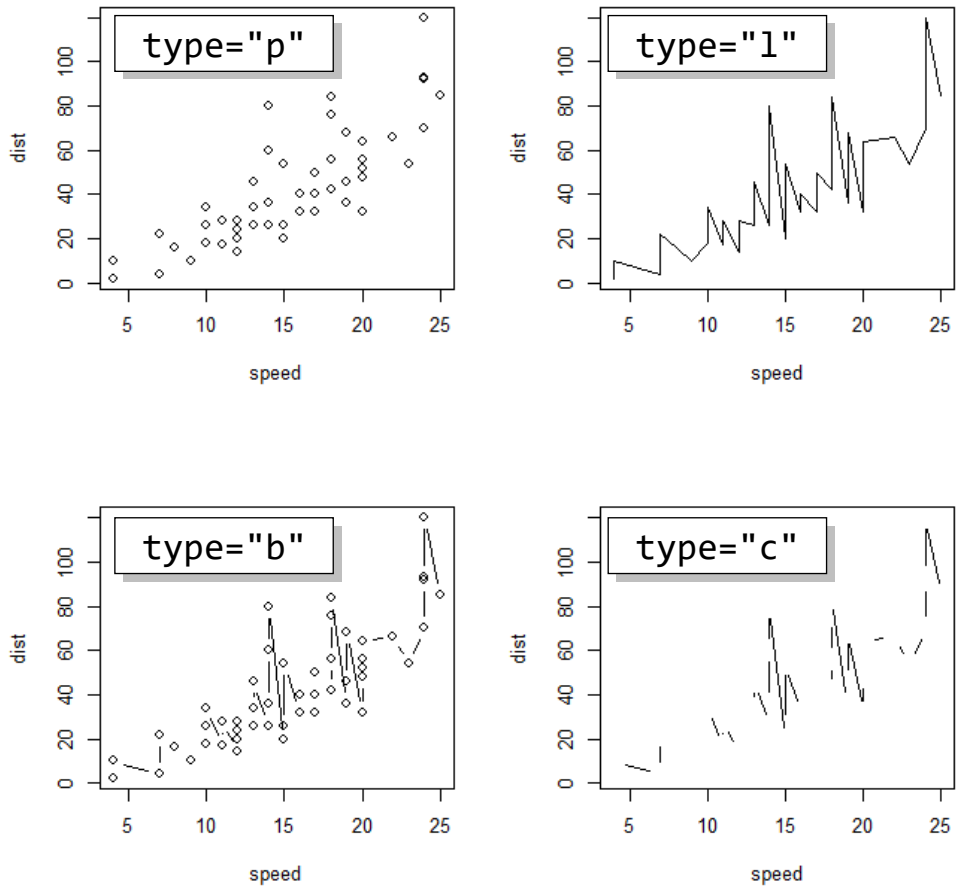
plot 関数は、引数で指定したオブジェクトの種類や構造に応じて、描く図を判断します。データフレームである cars を引数にしたとき、第一列を x 値、第二列を y 値、列名を軸の名称として散布図を描きます。描画される図は、上記と同様のものです。

```
> plot(cars)
```

今回の描画は、データセットのデータを点として描きましたが、そのイメージを type 引数で変更することができます。一般的に、type 引数は他の描画関数でも有効です。

type 引数の値	描画イメージ
p	点
l	線
b	点と線
c	b から点を除く
o	点と線
h	垂線
s	階段状(水平→垂直)
S	階段状(垂直→水平)

Type 引数を変更した例をいくつか掲載します。



■R を使用した円グラフの作成

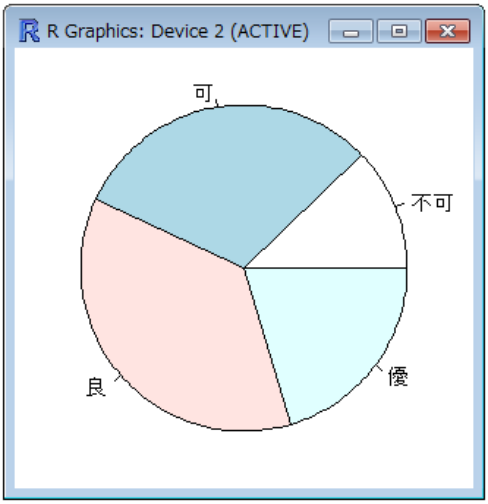
円グラフを描くには pie 関数を使用します。下記の例では、「ある学校のあるクラスにおける科目の成績の分布データ」を使用して、成績の割合を円グラフで描画しています。

【成績の分布データ】

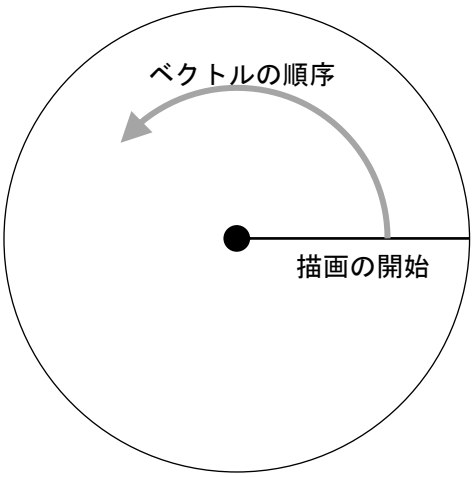
不可	可	良	優
15 名	38 名	45 名	25 名

それぞれの成績の割合を円グラフで描画するには、次のように行います。対象の要素(人数)の合計を 100%として、それぞれの要素の値の割合が描画されます。

```
>s <- c("不可"=15, "可"=38, "良"=45, "優"=25)
>par(mar=c(0,0,0,0), oma=c(0,0,0,0))
>pie(s)
```



↑描画結果



↑ベクトルの要素の順序と描画の関係
(反時計回り、0 度)

描画の順序を逆にしたり、開始角度を変更したりする場合、clockwise、init.angle といったオプションを使用します。描画結果はスクリプトを実行して確認してみてください。

```
> pie(s, clockwise=FALSE, init.angle=90, main="反時計回り 90 度")
```

■R を使用した箱ひげ図の作成

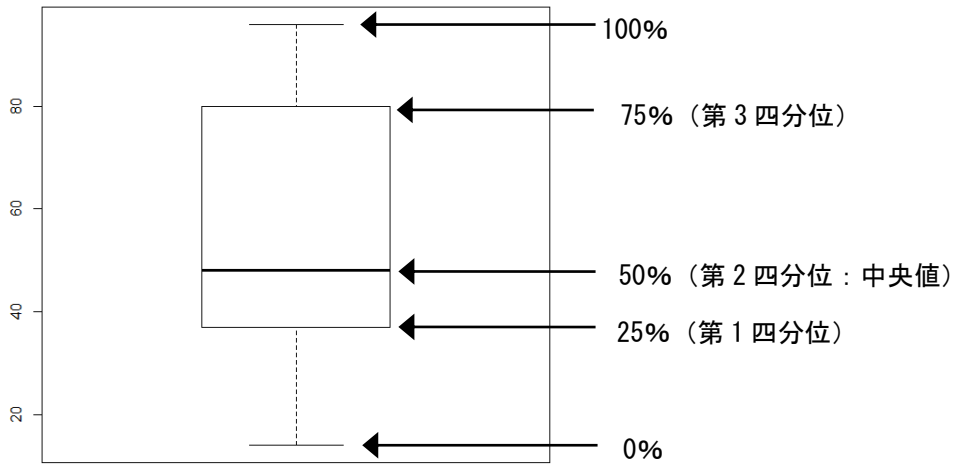
箱ひげ図を描くには boxplot 関数を使用します。下記の例では次のような「クラスごとのテストの成績データ(3 クラス分・score.csv)」を使用して、箱ひげ図を描画しています。

【クラスごとのテストの成績データ】

クラス	成績
1	80
～中略～	
2	40
～中略～	
3	28

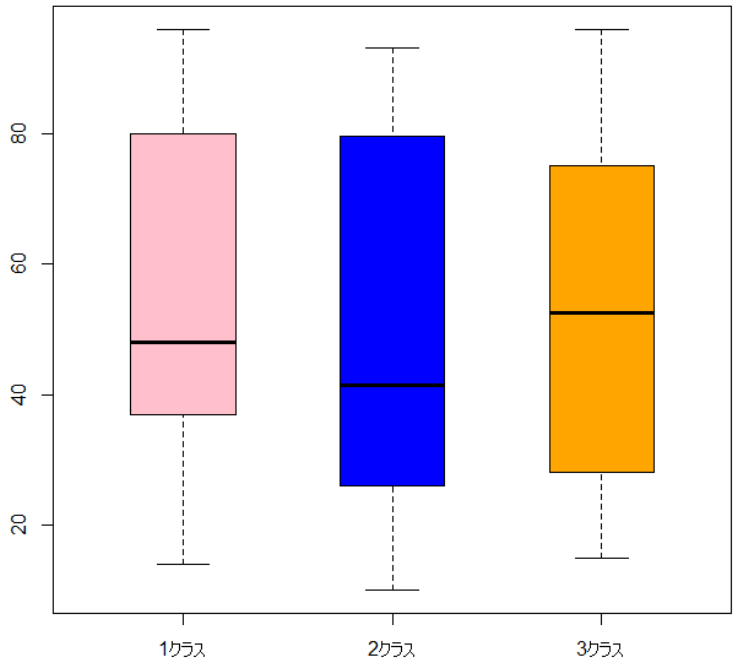
クラス"1"の成績についての箱ひげ図を描画するには、次のように行います。

```
> s <- read.table("score.csv", header=TRUE, sep=",")
> head(s, n=5)
  クラス 成績
1      1   80
2      1   62
3      1   58
4      1   47
5      1   36
> s2 <- as.vector(s[s[,1] == 1, 2])    # クラス"1"の成績のみを抽出
> quantile(s2)                         # 四分位点を確認
  0%  25%  50%  75% 100%
  14   37   48   80   96
> boxplot(s2)
```



クラスごとの箱ひげ図を並べて描画するには、次のように行います。

```
> boxplot(成績 ~ クラス, data=s, boxwex=0.5,
  names=c("1 クラス", "2 クラス", "3 クラス"), col=c("pink","blue","orange"))
```



3．5 グラフを見るときのポイント

グラフを見ることにより、データがどのような特徴をもっているかを代表値とは違う視点で理解し、どのようなアクションを起こしていくかを検討する必要があります。

グラフには様々なものがあり、活用方法もいろいろあります。本講義で紹介する統計解析手法とともによく使用する定型レポートは、**「ヒストグラム」**、**「折れ線グラフ」**、**「散布図」**です。

理想とするデータの特徴ごとの、「ヒストグラム」、「折れ線グラフ」、「散布図」の見るべきポイントは、以下のようになります。

理想とするデータの特徴	見るべきポイント
値が高いほど良い	【ヒストグラム】 **値が集中**しているか、**分散しているか**を確認する。分散している場合は、**値が変化する要因**を調査する。 【折れ線グラフ】【散布図】 **縦軸に値**、**横軸にそれに関連する要因**を配置し、横軸の値により、値が変化するかを確認する。
値が低いほど良い	【ヒストグラム】 **値が集中**しているか、**分散しているか**を確認する。分散している場合は、**値が変化する要因**を調査する。 【折れ線グラフ】【散布図】 **縦軸に値**、**横軸にそれに関連する要因**を配置し、横軸の値により、値が変化するかを確認する。
特定の値が望ましい	【ヒストグラム】 **値が理想とする値に集中**しているかを確認する。理想とする値から外れている場合は、その要因を調査する。 【折れ線グラフ】【散布図】 **縦軸に値**、**横軸にそれに関連する要因**を配置し、横軸の値により、値が変化しないかを確認する。

第3章演習問題 –代表値-

【ケース】 第 3 章～第 5 章の演習問題では、同様のケースを使用して簡易的な演習を行います。

Honda 書店.com は、IT 技術を中心とした書籍をネット販売する会社である。売上は年々上昇しており順調ではあるが、さらなる売上向上を実現するため、業務データ、ソーシャルデータ、アクセスログなどを分析して導き出した知見を活用していくことを検討している。

(1)IT 技術系の本の分類ごとの売上データを準備しました。各分類の代表値を求め、各分類の売上の特徴を分析してください。

分類ごとの売上データは下記ファイルに格納してあります。データを R に読み込み、分析して下さい。

分類ごとの売上データ ： c:¥kda¥text¥prac3.csv

●分類ごとの売上データ

日付	プログラム	NW	DB	OS	統計
2015/4/1	35	82	46	76	60
2015/4/2	36	82	45	75	63
2015/4/3	40	89	47	77	91
2015/4/4	32	69	43	73	6
…	…	…	…	…	…

●分類ごとの代表値

	プログラム	NW	DB	OS	統計
平均					
中央値					
合計					
最大					
最小					
件数					
分散					
標準偏差					

分類	考察
プログラム	
NW	
DB	
OS	
統計	

第4章　相関分析

４．１　相関分析とは
４．２　相関係数の算出

4．1 相関分析とは

相関分析

■相関分析とは
2つの項目間に関連があるかを調べる

■データの準備
関連を調べる値の組み合わせ

■分析の実施
cor関数を使用して分析可能

■相関分析とは

相関分析とは、2 つの項目間の関連の強さを調べる分析です。項目間の関連の強さを調べるために、相関係数などの関連の強さを表現する値を計算します。

■データの準備

連続した数値をもつ2つの項目の組み合わせを用意します。例えば、**「アクセス数×売上数」**、**「部署人数×消費電力」**などの組み合わせです。

アクセス数が増えるほど売上数が増える関係なのか、アクセス数が増えるほど売上数が減る関係なのか、ほとんど関係がないのかを調べることができます。

【サンプル】Web サイトアクセス数と売上数

アクセス数	売上数
9	4
30	5
11	11
…	…

【相関分析を実施するケース】

理想とするデータの特徴	分析結果の活用方法
値が高い(低い)ほど良い	・管理する値に影響の強い項目を探し出し、その項目に働きかけ、値を向上(低下)させる
特定の値が望ましい	・管理する値に影響の項目を探し出し、その項目に働きかけ、値を安定させる

【活用事例】

・つぶやきのキーワード数と株価との関係を調べ、株価予測に活用する。
・Web サイトのデザインと購買実績との関係を調べ、それをもとに Web サイトを改善し、
売上を向上する。
・ある商品とある商品の購買に関連があるかを調べ、関連の強い商品の組み合わせをリコメンドする。

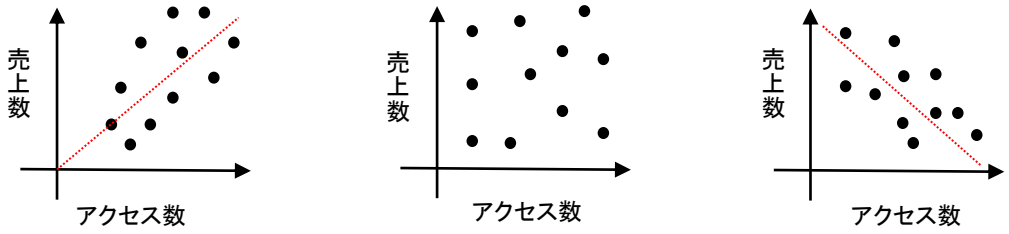
■分析の実施

量的変数(※)×量的変数の関係を調べるためには、**相関係数**を算出します。相関係数は必ず「－1～1」の値を取り、その値の解釈は、以下のようになります。

※量的変数:連続した値で表現される、計算のできる(計算結果に意味のある)変数のこと

0.3 ＜【相関係数】≦ 1 :正の相関あり
-0.3 ≦【相関係数】≦ 0.3 :相関なし
-1 ≦【相関係数】＜ -0.3 :負の相関あり

2 つの項目で散布図を描くと、以下のような特徴になります。



正の相関あり　　相関なし　　負の相関あり

相関係数は、2 つの項目間に直線的な関係があるかを調べる方法です。2 つの項目間に関係があったとしても、その関係が直線的ではないと相関係数は、0 付近の値になってしまいます。そのため、散布図で項目間にどのような関係があるかを把握し、算出された相関係数を適切に評価、適用する必要があります。

4．2 相関係数の算出

■相関係数の算出

「量的変数×量的変数」の項目間の関連の強さを調べるためには、相関係数を算出します。
相関係数の算出方法を以下に紹介します。

$$\text{相関係数} = \frac{\text{「項目1の偏差」と「項目2の偏差」の積の合計}}{\sqrt{\text{「項目1の偏差」の2乗の合計} \times \text{「項目2の偏差」の2乗の合計}}}$$

[計算例]

	アクセス数	販売数	②アクセス数の偏差	売上数の偏差	③アクセス数の偏差の2乗	売上数の偏差の2乗	④偏差の積
	9	4	-13	-2.9	169	8.3	37.6
	30	5	8	-1.9	64	3.6	-15.1
	11	11	-11	4.1	121	16.9	-45.2
	35	6	13	-0.9	169	0.8	-11.6
	9	2	-13	-4.9	169	23.9	63.6
	7	8	-15	1.1	225	1.2	-16.7
	4	4	-18	-2.9	324	8.3	52.0
	48	7	26	0.1	676	0.0	2.9
	45	15	23	8.1	529	65.8	186.6
①平均	22	6.9		⑤合計	2446	128.9	254

①各項目の平均値を算出します。
②各項目の偏差を求めます。偏差は、「個別値-平均値」です。
③各項目の偏差の値を2乗します。
④各項目の偏差の積を算出します。
⑤③と④の合計値を算出します。
⑥以下の式で相関係数を算出します。

$$\text{「アクセス数」と「販売数」の相関係数} = \frac{254}{\sqrt{2446 \times 128.9}} = 0.45$$

■R での分析例（1）：1対1の関係性

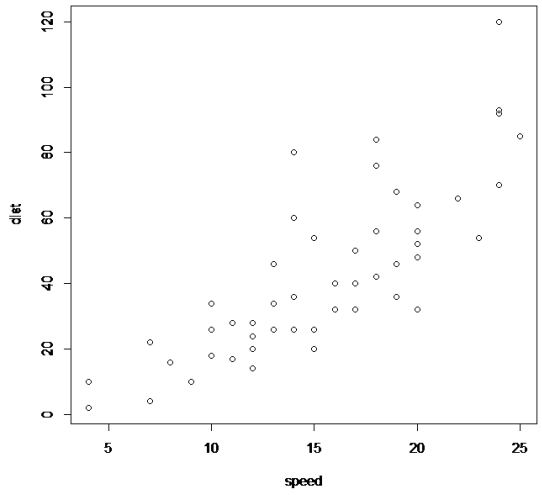
「車の速度と停止するまでの距離」の関係を示したデータセット cars について、speed（車の速度）と dist（停止するまでの距離）の相関係数を得るには、次のように行います。

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10

> cor(cars$dist,cars$speed,method="pearson")
[1] 0.8068949

> cor(cars)
          speed      dist
speed 1.0000000 0.8068949
dist  0.8068949 1.0000000
> plot(cars)  # 散布図を描画
```

cor 関数：第一引数、第二引数に調査するそれぞれのデータを指定します。

method 引数：相関係数を求めるアルゴリズムを指定します。デフォルト値は pearson（ピアソン）

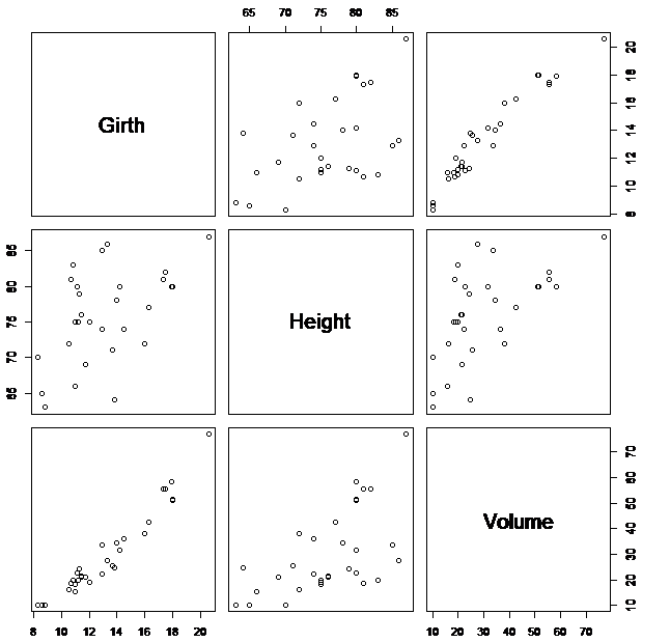
上記の例の場合、相関係数が約 0.80 という数値になっているので「speed と dist の間には強い正の相関関係がある」と言えます。



■R での分析例（2）：多対多の関係性

「アメリカ桜の周長（Girth）、高さ（Height）、容積（Volume）」の関係を示したデータセット trees において、周長（Girth）、高さ（Height）、容積（Volume）の相関係数を一括で算出し、その後散布図行列を作成します。

```
> head(trees)
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
6  10.8     83   19.7

> cor(trees)

           Girth    Height    Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000

> pairs(trees)   # 散布図行列を作成
```



第4章演習問題 –相関分析-

(1)分類「統計」の売上増加の原因を調べるために、「ビッグデータ」のつぶやき数、分類「統計」の売上数のデータを収集しました。この 2 項目間に関係があるかを分析してください。

上記データは下記ファイルに格納してあります。データを R に読み込み、分析して下さい。

c:¥kda¥text¥prac4.csv

●「ビッグデータ」のつぶやき数と分類「統計」の売上

日付	つぶやき数	売上
2011/4/1	1292	70
2011/4/2	1320	69
2011/4/3	1394	74
2011/4/4	1399	73
2011/4/5	1270	69
2011/4/6	1292	68
2011/4/7	1172	68
2011/4/8	1296	68
2011/4/9	1375	72
2011/4/10	1418	76
2011/4/11	1235	68
2011/4/12	1488	75
2011/4/13	1236	68
2011/4/14	1444	68
…	…	…

●つぶやき数と分類「統計」の売上との関係に対する考察

考察

(2)分類「統計」の売上が伸びている原因を分析するため、日付(Date)ごとの「ビッグデータに関連するつぶやき数(TW_Bigdata)」「ビッグデータ特集ページへのアクセス数(AC_Bigdata)」、「データベースに関連するつぶやき数(TW_DB)」「データベース特集ページへのアクセス数(AC_DB)」「アクセスした顧客の平均年齢(Age Sales)」を収集しました。それぞれの項目と分類「統計」の売上(Sales)との関係を分析してください。

上記データは下記ファイルに格納してあります。データを R に読み込み、分析して下さい。

c:¥kda¥text¥prac4-2.csv

●分類「統計」の売上と各項目の実績

Date	TW_Bigdata	AC_Bigdata	TW_DB	AC_DB	Age	Sales
2011/4/1	1292	1540	1563	1613	48	70
2011/4/2	1320	1399	1260	1310	49	69
2011/4/3	1394	1281	893	943	39	74
2011/4/4	1399	1338	1043	1093	44	73
2011/4/5	1270	1396	1254	1304	41	69
2011/4/6	1292	1408	1301	1351	40	68
2011/4/7	1172	1271	988	1038	43	68
・・・	・・・	・・・	・・・	・・・	・・・	・・・

●「ビッグデータ入門」と「統計入門」の購買に対する考察

項目	考察
つぶやき (ビッグデータ)	
アクセス数 (ビッグデータ)	
つぶやき (データベース)	
アクセス数 (データベース)	
アクセス 平均年齢	

第5章　回帰分析、判別分析

５．１　回帰分析とは
５．２　判別分析（ロジスティック回帰分析）とは

5．1 回帰分析とは

回帰分析

■回帰分析とは
値を予測する予測式を作成

■データの準備
予測する値と相関の強い値の組み合わせ

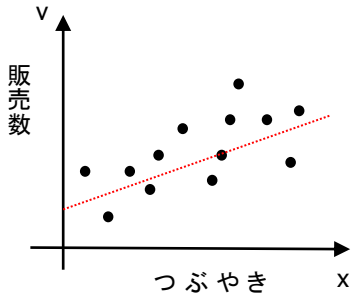
■分析の実施
lm関数を使用して分析可能

■回帰分析とは

回帰分析とは、ある指標(**目的変数**)と他の変数(**説明変数**)の関係を調べ、説明変数の値から目的変数の値を予測する式を作成する分析です。予測式を作成してもその予測式の精度が悪いと意味がありませんので、予測式とともに、**決定係数(R²値)**と呼ばれる値も算出します。決定係数(R²値)は、0 から1までの値を取り、値が大きいほど予測式の精度が高いと言えます。例えば、予測式は、以下のような式になります。

y = 0.013x + 4.4626　R² = 0.6095

y は販売数(目的変数)、x はつぶやき数(説明変数) 、R²は決定係数



説明変数が 1 つである回帰分析を**単回帰分析**、説明変数が複数ある回帰分析を**重回帰分析**といいます。また、回帰分析で使用される最も基本的なモデルは「Y=aX + B」という線形回帰となります。

【回帰分析を実施するケース】

理想とするデータの特徴	分析結果の活用方法
・値が高い(低い)ほど良い ・特定の値が望ましい	・予測式を作成し、目的変数の値を予測する。予測した値が望ましくない値である場合は、事前対策を実施する。 ・ある項目の値が、管理する値にどれくらい影響を与えるかを予測式から把握する。

【活用事例】

・つぶやきのキーワードの数と株価との関係を調べ、株価を予測する。

・店舗面積、周辺の世帯数、販売商品数から、新店舗の売上を予測する。

・つぶやき数、ブログアクセス数、CM 時間などから、購買数を予測する。

■データの準備

回帰分析を実施するためには、予測したい値(目的変数)とそれに関連する値(説明変数)の組み合わせを準備します。単回帰分析であれば、目的変数と1つの説明変数の値の組み合わせを準備します。重回帰分析であれば、目的変数と複数の説明変数の値の組み合わせを準備します。

説明変数は、**目的変数と相関の強い項目**を準備すると予測式の精度は高くなります。また、データ数が多いほど、分析の精度が高まるため、より多くのデータを準備します。

・単回帰分析用のデータ

目的変数と説明変数の組み合わせを用意します。

つぶやき数	販売数
228	4
87	5
492	11
58	6

・重回帰分析用のデータ

目的変数と複数の説明変数の組み合わせを用意します。説明変数が多いほど、様々な予測式を検証することができるので、より多くの説明変数を準備します。

つぶやき数	ブログ数	販売数
228	46	4
87	21	5
492	127	11
58	12	6

■R での単回帰分析の実施例

「車の速度と停止するまでの距離」の関係を示したデータセット cars について、speed を説明変数に、dist を目的変数にして回帰式を作成します。回帰分析を行う際は「lm 関数」を使用します。

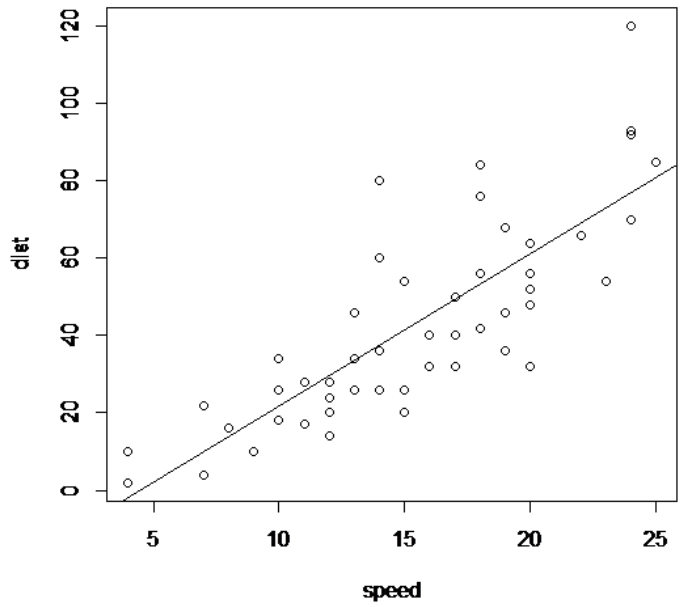
```
> res<-lm(dist~speed,data=cars)  # 変数 res に分析結果を代入
> summary(res)                   # summary 関数で内容を確認

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

分析結果より、下記内容が確認できます。
回帰式　:dist = speed × 3.9324 − 17.5791
決定係数 :0.6511

lm 関数:「目的変数~説明変数」といった形でデータを指定します。
data 引数:使用するデータを指定します。

plot 関数で散布図を表示し、abline 関数を使用して回帰直線を可視化することも可能です。

```
> plot(cars)

> abline(res)
```



■R での重回帰分析の実施例

「アメリカ桜の周長(Girth)、高さ(Height)、容積(Volume)」の関係を示したデータセット trees において、目的変数を容積(Volume)、説明変数を周長(Girth)と高さ(Height)、として重回帰分析を行います。

```
> res  <- lm(Volume ~ Girth + Height, data = trees)
> summary(res)

Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q
-6.4065 -2.6493 -0.2876  2.2003  8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16  ***
Height        0.3393     0.1302   2.607   0.0145  *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,    Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

分析結果より、下記内容が確認できます。
回帰式　:Volume = 4.7082 × Girth ＋ 0.3393 × Height ― 57.9877
補正済み決定係数　:0.9442

重回帰分析の場合、決定係数は Ajusted R-squared を参照します。

【注意事項】

重回帰分析を実施した場合、決定係数が高くても、精度が悪い場合があります。重回帰分析をおこなうと各説明変数に対し、P 値というものが算出されますが、この値が 0.05 以上の説明変数が存在すると、決定係数が高くても、精度が悪くなります。その場合は、その説明変数を外し、再実行してみてください。

5．2 判別分析（ロジスティック回帰分析）

判別分析（ロジスティック回帰分析）

■ロジスティック回帰分析とは
0～1の可能性を予測する予測式を作成

■データの準備
予測する値と相関の強い値の組み合わせ

■分析の実施
glm関数を使用して分析可能
（オプション指定の必要あり）

■ロジスティック回帰分析とは

判別分析を行うための分析手法にはいろいろなものが存在しますが、目的変数が「はい」「いいえ」や「1」「0」のような 2 値を持ち、複数の説明変数によって、その目的変数の発生確率を予測するものに、**ロジスティック回帰**(Logistic regression)があります。

活用事例としては「顧客の属性データを使用して、ある商品を購入するかどうかを予測する」「患者の属性データを使用して、特定の病気に罹患するかどうかを予測する」「既存顧客のデータを使用して、自社サービスから離反するかどうかを予測する」といったような形で、結果が 2 値かつ、その確率を予測したいようなケースが考えられます。

■データの準備

ロジスティック回帰分析を実施するためには、予測したい値(目的変数)とそれに関連する値(説明変数)の組み合わせを準備します。今回は「ある予備校での模擬試験の点数と部活動、性別のデータ(説明変数)と、そのデータの受験生が某大学に合格したかどうか(目的変数:0/1 の 2 値データ)のデータ」を用います(本分析例用の架空のデータです)。下記にデータの一部分を掲載します。

国語	数学	英語	理科	社会	部活動	性別	合否
69	71	61	73	55	運動系	男性	1
54	21	56	94	85	文科系	男性	0
60	49	27	61	82	文科系	男性	0
15	64	46	58	49	無	女性	0

■ R でのロジスティック回帰分析実施例

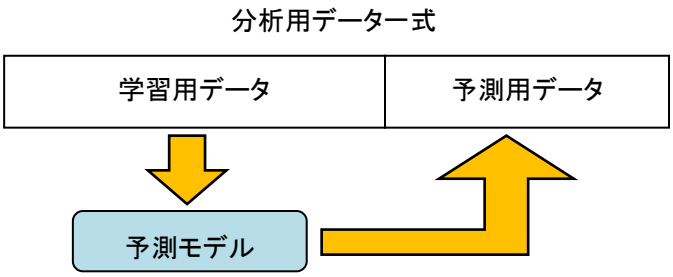
1)データの準備

次のようなテストの得点と合否結果のデータを扱います。合否結果を目的変数、その他データを説明変数として、その関係性を分析します。目的変数の値は「1(合格)」「0(不合格)」の値です(文字データも対象にすることが可能です)。

```
> judge <- read.table("judge.csv", header = T, sep = ",")
> head(judge)
  国語 数学 英語 理科 社会 部活動 性別 合否
1   69   71   61   73   55 運動系 男性    1
2   54   21   56   94   85 文科系 男性    0
3   60   49   27   61   82 文科系 男性    0
4   83   49   47   58   54 運動系 女性    0
5   35   17   58   25   20 文科系 女性    0
6   15   64   46   58   49     無 女性    0
> dim(judge)
[1] 200   8
```

2)学習用データと予測用データの作成

学習用のモデル(予測式)を作成し、その精度を評価する場合、存在するデータセットを**「学習用(モデル作成用)」のデータ(訓練データ)と「予測用(精度評価用)」のデータ(テストデータ)に分割する**必要があります。イメージは下記の通りです。



データの分割割合に決まりはありませんが、学習用:予測用の割合を「9:1」「8:2」「7:3」などにするのが一般的です。

今回は 200 件のデータがあるため、学習用データとして 150 件、予測用データとして 50 件のデータを使用します。

学習用データして、150 件のランダムな行番号を生成します。

```
> set.seed(4); s <- sample(1:200, size = 150)
```

テキストでの記載結果と実行結果をそろえるため、下記を行っています。

・set.seed(X)で、ランダムな数値生成のシード値を固定

⇒ シード値を他の値に設定すると結果が変わります。

・sample 関数を使用して、1～200 の数値の間から 150 件分のランダムな数値を生成し、s に代入を行う

上記により、訓練データは judge[s,] 、テストデータは judge[-s,] となります。

また、訓練データの内容によっては、適切なモデルが構築できない場合(予測が不正確)があります。そのようなときには訓練データを再作成します。

3）ロジスティック回帰分析の実行

R でロジスティック回帰を行うには、glm 関数を利用し、family オプションに binomial を指定します。その他、関数の第 1 引数には目的変数と説明変数の関係を指定し、data 引数に訓練データを指定します。関数の戻り値によりモデルが取得できます。

```
> judge.glm <- glm(合否 ~ 国語 + 数学 + 英語 + 理科 + 社会 + 部活動 + 性別,
                   family = binomial,
                   data = judge[s, ])
> step(judge.glm)
Start:  AIC=58.06
合否 ~ 国語 + 数学 + 英語 + 理科 + 社会 + 部活動 + 性別

【中略】
Step:  AIC=54.53
合否 ~ 国語 + 数学 + 英語 + 社会 + 部活動

         Df Deviance     AIC
<none>        40.527  54.527
- 部活動  2   45.916  55.916
- 社会    1   45.248  57.248
- 英語    1   74.196  86.196
- 国語    1   87.520  99.520
- 数学    1   91.405 103.405

Call:  glm(formula = 合否 ~ 国語 + 数学 + 英語 + 社会 + 部活動, family = binomial,
    data = judge[s, ])

Coefficients:
 (Intercept)          国語          数学          英語          社会
   -35.44318       0.16237       0.14827       0.18125       0.04989
部活動文科系      部活動無
     0.72985      -5.10533

Degrees of Freedom: 149 Total (i.e. Null);  143 Residual
Null Deviance:      165.3
Residual Deviance: 40.53        AIC: 54.53
```

前頁では、目的変数を「合否」、説明変数に残りすべての変数を設定して、予測モデルを「judge.glm」に格納しています。その後、step 関数を使用してステップワイズ法を実施し、目的変数に影響を与える項目（説明変数）についての示唆を得ています。

今回のケースの場合、「合否を予測する場合、 国語 + 数学 + 英語 + 社会 + 部活動 で予測するのが良さそうだ」という結果を得ることができました（今回の学習用データを使用した場合、上記の結果になったということなので、データを変えると別のモデルになる可能性もあります。また、今回使用しているデータはあくまでもダミーで作成したデータとなります）。

【参考】ステップワイズ法とは

モデル作成の際、使用可能な説明変数が複数存在する場合は、説明変数の取捨選択を行う必要があります。理想としては、出来上がるモデルと関係性の高い説明変数を把握しておき、モデル作成に使用していくのがベストですが、「説明変数の数が多い」「どの説明変数を使えばいいのかわからない」という場合も考えられます。

その際、説明変数を増減させつつ出来上がったモデルの精度やモデル判断指標を確認していき、「どの説明変数の組み合わせでモデルを作成するのが適切か」を確認するために使用するのがステップワイズ法です。

R では step 関数を使用することにより、AIC をベースとしたステップワイズ法を行うことが可能です。

【参考】AIC とは

AIC(Akaike Information Criterion : 赤池情報量規準)とは、モデルの当て嵌まり具合を測る指標で、算出された AIC が小さいモデルほど良いモデルと考えることができます。AIC ではパラメータ数の少ないモデルを良いモデルと判断するため、「できるだけ少ない説明変数を用いてデータを適切に予測する」モデルが AIC 的な観点からすると良いモデルということになります。

「国語 + 数学 + 英語 + 社会 + 部活動」の説明変数で再度予測モデルを作成し、summary 関数でモデルの詳細を確認します。

```
> judge.glm <- glm(合否 ~ 国語 + 数学 + 英語 + 社会 + 部活動 , family = binomial,
data = judge[s, ])
> summary(judge.glm)

Call:
glm(formula = 合否 ~ 国語 + 数学 + 英語 + 社会 + 部活動, family = binomial,
    data = judge[s, ])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.05968  -0.08442  -0.00432  -0.00008   1.90658

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -35.44318    9.22640  -3.841 0.000122 ***
国語           0.16237    0.04446   3.652 0.000260 ***
数学           0.14827    0.03689   4.019 5.83e-05 ***
英語           0.18125    0.05328   3.402 0.000669 ***
社会           0.04989    0.02609   1.912 0.055861 .
部活動文科系   0.72985    1.01165   0.721 0.470634
部活動無      -5.10533    2.69163  -1.897 0.057862 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 165.324  on 149  degrees of freedom
Residual deviance:  40.527  on 143  degrees of freedom
AIC: 54.527

Number of Fisher Scoring iterations: 9
```

judge.glm の cofficients で係数を確認することができますが、ここで出力されている値は対数オッズの値となるため、exp 関数で対数を外します（そのままだと見づらいため、round 関数で小数点第 2 位に丸めた結果を表示しています）。

```
> judge.glm$coefficients
 (Intercept)        国語        数学        英語        社会 部活動文科系      部活動無
-35.44318029  0.16236733  0.14826928  0.18125389  0.04988775  0.72985454 -5.10533000

> round(exp(judge.glm$coefficients),2)
 (Intercept)        国語        数学        英語        社会 部活動文科系      部活動無
        0.00        1.18        1.16        1.20        1.05        2.07        0.01
```

上記より、国語、数学、英語、社会の点数が上がると合格の確率が上がり、部活動が文科系の場合合格率が上がり、部活動無しの場合は合格率が下がっている、という結果が確認できます。

上記のモデルを使って、予測用のデータの合否を判定します。予測を行う場合、predict 関数を使用します。引数としてモデル(judge.glm)と予測用データ(newdata)を指定し、type を response に指定することにより、確率で結果を受け取ることができます。

ここでは、合格確率が 50%以上を合格、49%以下を不合格として簡易的に表示を行うため、round 関数を使用し、合格(1)と不合格(0)の予測結果を出力しています。

```
> (res <- round(predict(judge.glm, newdata = judge[-s,], type="response")))
  3   5  11  16  28  33  35  42  47  48  49  50  56  57  60  61  64  65
  0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   1   1   1
 70  77  83  88  93  94 103 106 109 111 116 121 125 131 132 140 145 147
  0   0   0   0   0   0   1   0   0   0   0   0   0   1   0   1   1   0
159 163 164 165 166 169 171 172 173 174 181 182 186 195
  0   1   1   0   0   0   1   0   1   0   0   0   0   0
```

4)予測結果の確認

予測用データ(judge[-s,])に、前頁で予測した res の結果を列として追加し、結果を確認します。

```
> judge.ALL <- transform(judge[-s,],PREDICT=res)
> head(judge.ALL)
   国語 数学 英語 理科 社会 部活動 性別 合否 PREDICT
3    60   49   27   61   82 文科系 男性    0       0
5    35   17   58   25   20 文科系 女性    0       0
11   72   56   76   44   68 文科系 男性    1       1
16   60   71   81   20   67 運動系 女性    1       1
28   36   50   86   41   59 運動系 女性    0       0
33   70   56   62   87   36     無 女性    1       0
```

上記の合否、PREDICT 列を比較することによって、正解率を求めることも可能ですが、ここではもう一歩踏み込んで、混同行列を作成します。

下記 1 行目では、table 関数を使用して、合否(1:合格、0:不合格)と予測結果(1:合格、0:不合格)のクロス集計表を作成し、2行目、3行目で結果をわかりやすくするため、0,1を入れ替えて表示を行っています。

```
> (cm <- table(judge.ALL$PREDICT,judge.ALL$合否)) # 列:正解値 行:予測値
> (cm2 <- cm[,c(2,1)])     # 列入れ替え
> (cm3 <- cm2[c(2,1),])    # 行入れ替え


    1  0
  1 11  2
  0  3 34
```

5)混同行列を使用した正解率、各種指標の確認

混同行列とは、分析モデルの精度や各種指標を計算するために**「分析モデルを使用した予測結果と実際の正解の一致、不一致をカウントして表にまとめたもの」**のことです。2 値分類の場合は下記のような形で表します。

		正解値	
		Positive	Negative
予測値	Positive	True-Positive (TP)	False-Positive (FP)
	Negative	False-Negative (FN)	True-Negative (TN)

Positive:検知対象、異常、不良品などを表すことが多い

Negative:検知対象外、正常、良品などを表すことが多い

※正解値、予測値の行列が逆に記載されるケースもあります。

※案件によっては Positive、Negative が上記と逆になるケースもあります。

このとき、下記のような指標を算出することができます。

・正解率、精度(Accuracy) ⇒ (TP + TN) / (TP + FP + FN + TN)

⇒ 全データにおける正解率を表します。

・再現率(Recall) ⇒ TP / (TP + FN)

⇒ 実際に Positive であるデータ中、Positive と予測できたものの割合

⇒ 取りこぼしを避けたい場合にこの指標を重視します。

・適合率(Precision) ⇒ TP / (TP + FP)

⇒ Positive と予測したデータ中、実際に Positive だったものの割合

⇒ 誤検知を避けたいときにこの指標を重視します。

・F 値(F measure) ⇒ (2 × 再現率 × 適合率) / (再現率+適合率)

⇒ 再現率と適合率の調和平均(再現率、適合率の総合的な指標)

指標としては、正解率が非常に理解しやすいですが、「取りこぼしを避けたい」「誤検知を避けたい」などの要件に応じて、再現率や適合率などを一緒に確認をしていくことになります。また、再現率と適合率に関しては、トレードオフの関係にあるため、要件に応じて適切な指標を選択する必要があります。

今回のケースの場合、計算結果は下記の通りになります。

		正解	
		1	0
予測	1	TP 11	FP 2
	0	FN 3	TN 34

・正解率、精度(Accuracy) ⇒ (TP + TN) / (TP + FP + FN + TN)

⇒ 45 / 50 = 0.9

・再現率(Recall) ⇒ TP / (TP + FN)

⇒ 11 / (11 + 3) = 0.786

・適合率(Precision) ⇒ TP / (TP + FP)

⇒ 11 / (11 + 2) = 0.846

・F 値(F measure) ⇒ (2 × 再現率 × 適合率) / (再現率＋適合率)

⇒ (2 * 0.786 * 0.846) / (0.786 + 0.846) = 0.815

上記を計算した結果を以下に記載します。

```
> cm3.accuracy <- sum(diag(cm3)) / sum(cm3) # 正解率
> cm3.recall <- cm3[1,1] / (cm3[1,1] + cm3[2,1]) # 再現率
> cm3.precision <- cm3[1,1] / (cm3[1,1] + cm3[1,2]) # 適合率
> cm3.Fnum <- (2 * cm3.recall * cm3.precision) /
              (cm3.recall + cm3.precision)      # F 値
> (cm3.resall <- data.frame("指標" =c("正解率","再現率","適合率","F 値"),
              "値"=c(cm3.accuracy,cm3.recall,cm3.precision,cm3.Fnum)))
    指標        値
1 正解率 0.9000000
2 再現率 0.7857143
3 適合率 0.8461538
4    F 値 0.8148148
```

第5章演習問題 -回帰分析-

(1)「ビッグデータ」のつぶやき数と分類「統計」の売上の関係が強いということで、その関係を利用し、分類「統計」の売上予測をすることになりました。実績値をもとに回帰分析をおこない、分類「統計」の売上予測式を作成してください。

上記データは下記ファイルに格納してあります。データを R に読み込み、分析して下さい。

c:¥kda¥text¥prac4.csv

●「ビッグデータ」のつぶやき数と分類「統計」の売上

日付	つぶやき数	売上
2011/4/1	1292	70
2011/4/2	1320	69
2011/4/3	1394	74
2011/4/4	1399	73
2011/4/5	1270	69
2011/4/6	1292	68
2011/4/7	1172	68
2011/4/8	1296	68
・・・	・・・	・・・

●分類「統計」の売上予測式

売上予測式

●回帰分析の考察

考察

(2)分類「統計」の売上とそれらに関連があると思われる項目を利用し、分類「統計」の売上予測をすることになりました。実績値をもとに回帰分析をおこない、分類「統計」の売上予測式を作成してください。

上記データは下記ファイルに格納してあります。データを R に読み込み、分析して下さい。

c:¥kda¥text¥prac4-2.csv

●分類「統計」の売上と関連項目の実績

日付	つぶやき（ビッグデータ）	アクセス数（ビッグデータ）	つぶやき（データベース）	アクセス数（データベース）	アクセス平均年齢	売上
2011/4/1	1292	1540	1563	1613	48	70
2011/4/2	1320	1399	1260	1310	49	69
2011/4/3	1394	1281	893	943	39	74
2011/4/4	1399	1338	1043	1093	44	73
2011/4/5	1270	1396	1254	1304	41	69
2011/4/6	1292	1408	1301	1351	40	68
2011/4/7	1172	1271	988	1038	43	68
・・・	・・・	・・・	・・・	・・・	・・・	・・・

●分類「統計」の売上予測式

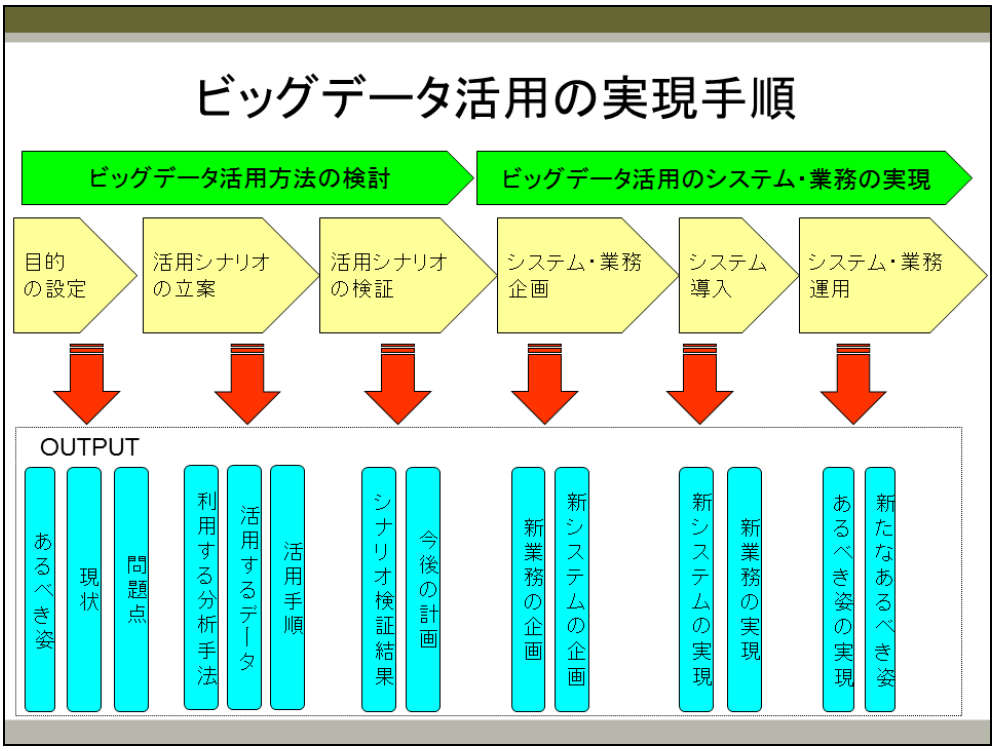
売上予測式

●回帰分析の考察

考察

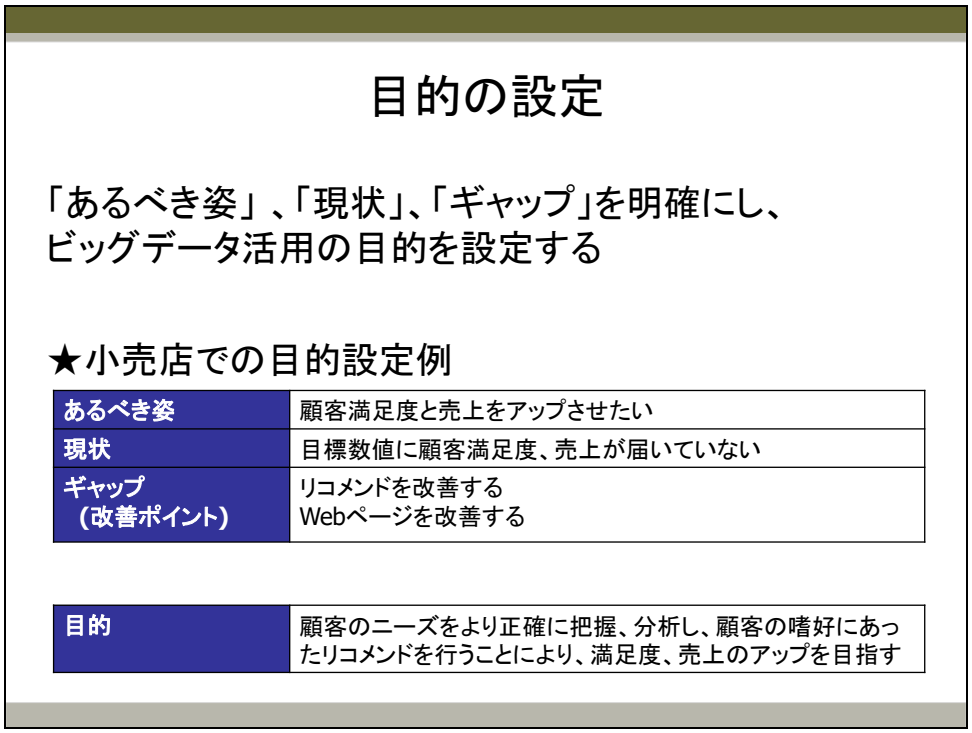
第6章　データ分析による課題解決
【演習】

6．1 ビッグデータ活用の実現手順とは



現在、ビッグデータ活用への期待が高まっていますが、「データの活用方法が分らない」、「対象とするデータが分らない」、「活用の方針が決まらない」などの課題を抱える企業が増えています。ここでは、ビッグデータ活用を実現するための手順を紹介します。

6．2 目的の設定



ビッグデータの活用は、**業務や生活を改善するためにおこなう手段**であり、**ビッグデータの活用自体が目的ではありません**。そのため、ビッグデータの活用方法を検討する前に、現在の業務や生活において、**何を改善すべきか**を考えます。改善すべきことを考えるためには、**「あるべき姿（TO BE）」**と**「現状（AS IS）」**を明確にし、そのギャップは何かを検討するとよいでしょう。そのギャップが改善すべき事項になります。改善すべき事項が複数挙げられたら、その中から実際に取り組む事項を決定します。それが、ビッグデータ活用の目的となります。

改善を実現するための方法は、ビッグデータ活用だけではありません。その他の方法も検討し、**ビッグデータ活用が改善に役立つと判断した時**に、ビッグデータを活用してください。

■ビッグデータ活用の良い例

目的：顧客に適切な商品をリコメンドし、顧客満足度と売上の向上を実現したい

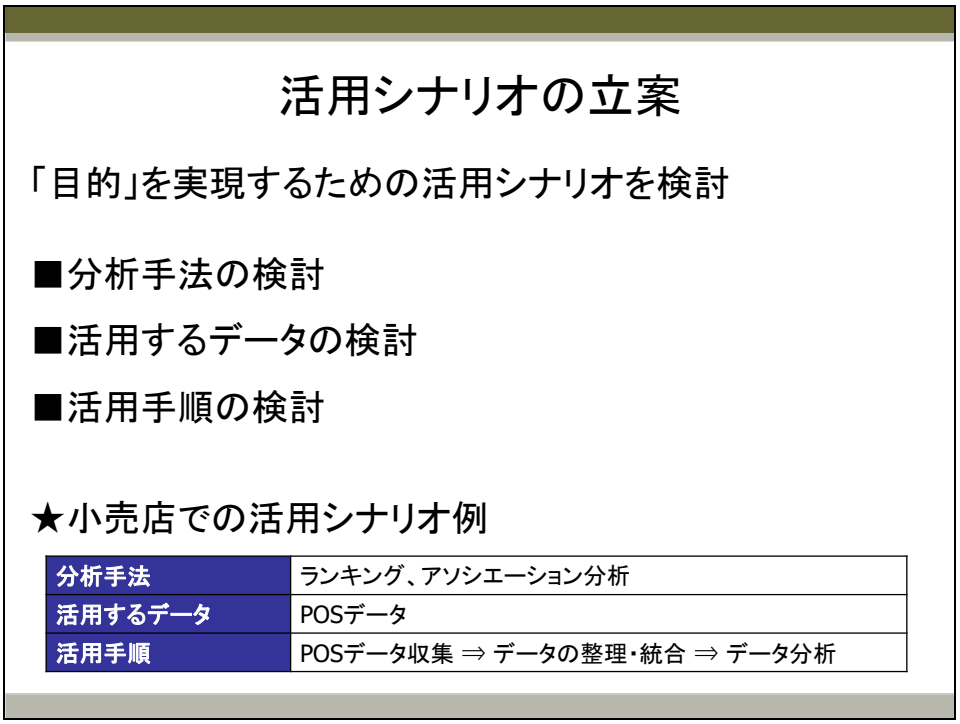
結果：ビッグデータを活用し、顧客と商品の相関を分析して顧客の嗜好を導出し、それに基づいて顧客の嗜好に合った商品をリコメンドしたところ、売上が向上した。

■ビッグデータ活用の悪い例

目的：ビッグデータを分析し、湿度と商品売上の相関を導出したい

結果：湿度の高い日には商品 A がよく売れ、湿度の低い日には商品 B がよく売れるということが分かったが、湿度に基づいて陳列商品を切り換えるしくみがないため、役に立たなかった。

6．3 活用シナリオの立案



改善すべき事項が決まりましたら、それを実現するための**ビッグデータの活用シナリオ**を検討します。活用シナリオを検討する際には、「分析手法」「活用するデータ」「活用手順」を検討します。

■分析手法

ビッグデータを分析してどのような知見を導き出すかを検討します。利用する**分析手法を検討するためには、分析手法の種類とそれぞれの特徴を理解しておく必要があります。**

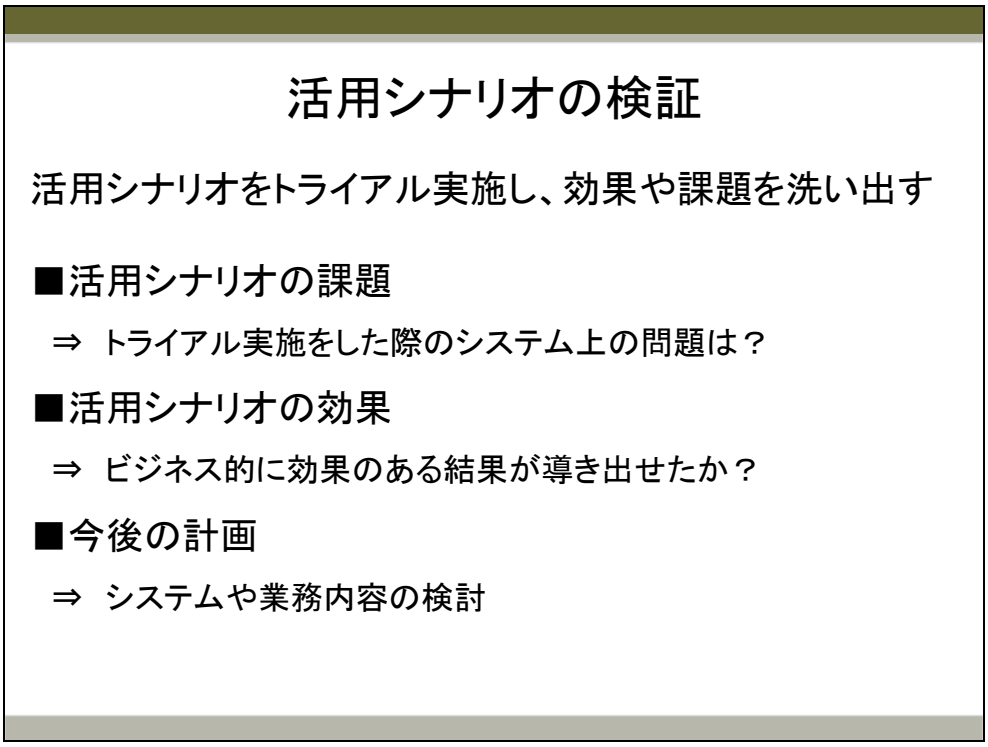
■活用するデータ

利用する分析手法ごとに必要なデータ形式でデータを準備する必要があります。データ分析は、活用するデータの種類が多いほど多くの知見が導出でき、データ数が多いほど分析結果の精度が高まる傾向があります。そのため、可能な限り、様々な種類、多くのデータを活用することをおすすめします。

■活用手順

ビッグデータを活用するためには、おもに「生成」「転送」「格納」「整理」「分析・対処」の手順が必要です。それぞれの手順を具体的にどのように実現するかを検討します。

6．4 活用シナリオの検証



活用シナリオが決まりましたら、そのシナリオをトライアル実施し、効果や課題を洗い出します。

■活用シナリオの課題の検証

活用シナリオを実践した際に、発生した問題点を列挙します。「生成」「転送」「格納」「整理」「分析・対処」の各手順で問題が発生した場合、本番系に移行する前のタイミングで各問題を解決しておく必要があります。

■活用シナリオの効果の検証

トライアル分析により、活用目的を実現するために効果のある分析結果を導くことができたかどうかを評価します。評価のポイントは、分析結果の内容と精度になります。

分析結果の内容が価値ある結果ではなかった場合は、分析手法、もしくは活用するデータを変更し、新しい結果を導き出す必要があります。また、精度に問題がある場合は、活用するデータの精度を上げるか、活用するデータの件数を増やす必要があります。

■今後の計画

活用シナリオを実践して得られた効果と課題をもとに、今後のビッグデータ活用の取り組みへの計画を検討します。ビッグデータ活用の実現を決断した場合は、課題の解決策を検討し、新しいシステムや業務内容を検討していきます。しかし、活用シナリオの実践により十分な効果が得られなかった場合は、目的や活用シナリオの見直し、もしくは撤退を検討します。

6．5 データ分析による課題解決演習

データ分析による課題解決演習

●分析報告書の作成
　ーユースケースに応じたデータ分析＆報告書の作成を行う

●使用可能なデータ
　ー日本の訪日外国人観光客に関するデータ
　ーある会社のエンゲージメントスコアに関するデータ
　ーあるECサイトにおけるユーザー属性データ

●プレゼンテーションの実施（5分＋QA3分）

本章では、ユースケースに応じたデータ分析を行い、クライアントに対して分析報告書を作成し、プレゼンテーションという形で発表を行います。

使用できるデータと、データの格納場所は下記の通りです。

各フォルダに、それぞれのデータについての説明と、分析の参考用のスクリプト（**そのままでは発表用としては使えないため、編集して使用してください**）が格納してありますので、内容確認の上、どれか1つの課題を選択してください。

- **日本の訪日外国人観光客に関するデータ**
  ⇒ C:¥kda¥Exercises¥JP
- **ある会社のエンゲージメントスコアに関するデータ**
  ⇒ C:¥kda¥Exercises¥ENG
- **ある EC サイトにおけるユーザー属性データ**
  ⇒ C:¥kda¥Exercises¥EC

それぞれのデータにおいてクライアントを想定し、分析目的を設定した上でデータ分析を行い、クライアントに対して状況改善のためのアクションプランに関する報告書を作成の上、プレゼンテーション（3～5 分間＋質疑応答 3 分間）を行ってください。

6．6 今回の報告書に含めるべきこと

今回の報告書に含めるべきこと

- ●報告書のサマリー
- ●分析目的
- ●分析手法
- ●使用したデータ
- ●分析結果
- ●得られた知見 → アクションプラン

今回の分析報告書に最低限含める事項は上記のスライドに記載した項目です。

必要に応じて、データの加工方法、検証方法や今後の課題などを記載してください。

6．7 報告書作成の際の留意点

報告書作成の際の留意点

●報告書のサマリーを最初につける

●報告書には「分析から得られた知見」と「アクションプラン」を盛り込む

●報告書は読み手の立場に立って作成する

報告書を作成する際は、読み手の立場に立って、下記に留意して作成しましょう。

1．報告書のサマリーを必ず最初につけること

・報告を受ける側の負担を軽減するため

・ファーストビューで大まかな内容や取るべきアクション、理由を把握するため

⇒ その後に理由や経緯を説明すること

2．分析報告書には「分析から得られた知見」と「アクションプラン」を盛り込むこと

・分析結果が出たからどうなのか？何がわかるのか？

⇒ 「分析結果がこうでました」だけでは意味がない

⇒ アクションプランを提示しなければ価値が創出できない

⇒ それを使って何ができるのか？

3．分析報告書は読み手の立場に立って作成すること

・その書き方で読み手に伝わるか？分析結果を張り付けただけになっていないか？

⇒ 分析手法や分析結果が「相手に」わかりやすく伝わるように記載する

⇒ 自分達がわかれば良い、ではダメ

・ ⇒ 読み手の立場にたって、必要最小限の内容を記載

⇒ 導き出した結論を報告すること

⇒ 分析結果だけでは「だから何？」が伝わらない

6．8 データ分析前の作業

データ分析前の作業

実際のデータ分析を行う場合は、
必要に応じて下記のような作業を行う

●データの結合、抽出
●データクレンジング
●データの変換
●変数の作成、追加

今回用意したデータセットは、元々存在するデータから抽出、加工などを行い、分析しやすい形に変換済みのデータとなります。そのため、実際にデータ分析を行う際は、

・データの結合、抽出

複数のテーブルを結合する、必要となるデータ部分のみを抜き出す作業

・データクレンジング

文字コード統一、名寄せ、重複値や欠損値、外れ値の処理などの作業

・データの変換

対数変換などのデータ変換、カテゴリ変数のダミー変数化、データの2値化など

・変数の作成、追加

例:「前日との気温差」列の作成、生年月日と購入日から「購入時の年齢」列作成、
来店者のポイントカードのデータから、その月の来店回数など

などを行う必要があります。

なお、どのような分析を行うかによって、上記の中でも必要な対応事項、必要のない事項などが出てくるため、「まずは分析の目的をしっかりと立てる」ところから始めると、無駄が出づらくなるでしょう。

【参考1】データクレンジング

データの中に不適切な内容のものが含まれていると、R での処理にエラーが発生したり、結果が誤ったものになったりします。このようなデータを適切な形式に変換する操作をデータクレンジングと呼びます。下記に代表的なデータクレンジング方法を記載します。

・重複値
・欠損値
・外れ値
・個数の不揃い

■重複値

ベクトル内の要素における重複値の削除は、次のように行います。

```
>a <- c(1, 2, 3, 3, 4, 5, 5, 5)
>(a2 <- unique(a))          # 重複した要素を削除する
[1] 1 2 3 4 5
```

データフレームにおける重複した行の削除は、次のように行います。

```
>(b <- data.frame("店舗"=c("A", "B", "A"), "年月"=c("2014/04", "2014/04",
"2014/04"), "売上"=c(100, 200, 100)))
  店舗    年月 売上
1    A 2014/04  100
2    B 2014/04  200
3    A 2014/04  100
>(b2 <- unique(b))            # 重複した行を削除する
  店舗    年月 売上
1    A 2014/04  100
2    B 2014/04  200
```

行列における重複した行の削除は、次のように行います。

```
>(c <- matrix(c(1, 2, 1, 3, 4, 3, 5, 6, 5), ncol=3, nrow=3))
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
[3,]    1    3    5
>(c2 <- unique(c))              # 重複した行を削除する
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

行列における重複した列の削除は、いったん行列を転置してから、前述の方法を用いて行います。

```
>(d <- matrix(c(1, 2, 3, 4, 5, 6, 1, 2, 3), ncol=3, nrow=3))
[,1] [,2] [,3]
[1,]    1    4    1
[2,]    2    5    2
[3,]    3    6    3
>(d2 <- t(d))                    # 行列を転置する
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    1    2    3
>(d3 <- unique(d2))              # 重複した行を削除する
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
>(d4 <- t(d3))          # 行列を転置する
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

■欠損値

ベクトル内の欠損値(NA)を削除するには、次のように行います。

```
>e <- c(1, 2, NA, 3, 4, NA, 5)
>(e2 <- e[!is.na(e)])          # 欠損値を削除する
[1] 1 2 3 4 5
```

データフレームにおいて、欠損値を含む行を削除するには、次のように行います。行列に対しても同じ方法で処理できます。

```
>(f <- data.frame("店舗"=c("A", "B", "C"), "年月"=c("2014/04", "2014/04",
"2014/04"), "売上"=c(100, NA, 200)))
  店舗    年月 売上
1    A 2014/04  100
2    B 2014/04   NA
3    C 2014/04  200
>(f2 <- na.omit(f))          # 欠損値を含む行を削除する
  店舗    年月 売上
1    A 2014/04  100
3    C 2014/04  200
```

ベクトル内の欠損値を特定の値で補完するには、次のように行います。特定の値としては、0 や平均値、最小値、最大値などの選択肢があります。この後に続く処理内容に応じて適切な値を選択します。

```
>g <- c(1, 2, NA, 3, 4, NA, 5)
>g[is.na(g)] <- mean(g[!is.na(g)])     # NAに平均値を代入
>g
[1] 1 2 3 3 4 3 5
```

データフレーム内の欠損値を特定の値で補完するには、次のように行います。行列に対しても同じ方法で処理できます。

```
>(h <- data.frame("店舗"=c("A", "B", "C"), "年月"=c("2014/04", "2014/04",
"2014/04"), "売上"=c(100, NA, 200)))
  店舗    年月 売上
1    A 2014/04  100
2    B 2014/04   NA
3    C 2014/04  200
>h[is.na(h)] <- 0            # NAに0を代入
>h
  店舗    年月 売上
1    A 2014/04  100
2    B 2014/04    0
3    C 2014/04  200
```

データフレームの列ごとに異なる値で補完するには、次のように行います。なお、因子である要素の上書きはエラーになることに注意します。

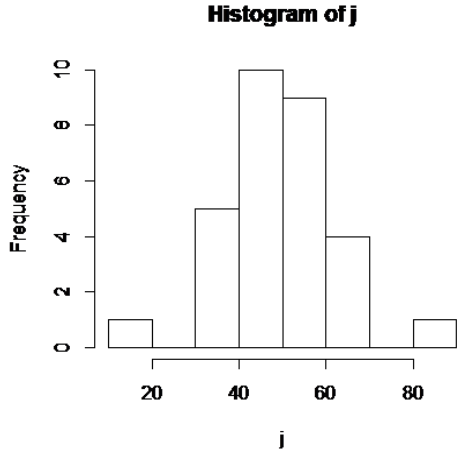
```
>(i <- data.frame("店舗"=c("A", "B", "C"), "年月"=c(NA, "2014/04", "2014/04"),
"売上"=c(100, NA, 200)))
  店舗    年月 売上
1    A <NA>  100
2    B 2014/04   NA
3    C 2014/04  200
>i[,2] <- as.character(i[, 2]) #「年月」列の因子を取り除く
>i[is.na(i[ ,2]), 2] <- "2014/01"       #「年月」列のNAを変換する
>i[is.na(i[ ,3]), 3] <- 0                #「売上」列のNAを変換する
>i
  店舗    年月 売上
1    A 2014/01  100
2    B 2014/04    0
3    C 2014/04  200
```

■外れ値

平均値などを算出する際、外れ値の存在が最適値を求めるための妨げになる場合があります。外れ値の判定は、標本の値がどのように分布するかにより異なります。

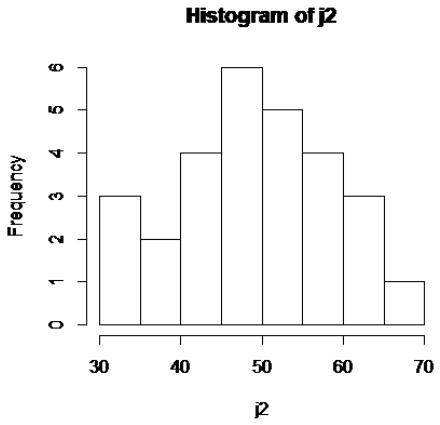
30 名の試験の得点をヒストグラムで描画します。

```
>j <- c(37, 59, 61, 41, 53, 50, 42, 63, 82, 52, 58, 59, 32, 46, 52, 51, 35,
45, 62, 53, 12, 41, 36, 48, 49, 69, 48, 57, 31, 49)
>hist(j)          # ヒストグラムの描画
```



試験の得点が正規分布に従うものと仮定して、平均値±標準偏差*2 の範囲を超えるものを外れ値と見なします（データやケースによるため、常にこれでいいということではありません）。外れ値を削除して、ヒストグラムを描画します。

```
>j2 <- j[j >= mean(j) –2*sd(j) & j <= mean(j) + 2*sd(j)]
>hist(j2)
```



■個数の不揃い

次のように、各行の項目数が一定でないデータの処理を考えます。

```
A, 85, 82, 44,  9, 74, 52, 64, 45, 11, 90
B, 11, 54, 89, 43, 68, 16, 65, 50
C, 50, 71, 93, 16, 34, 62,   ,  6, 18, 70, 53
D, 14, 36, 35, 52, 26, 76,  5
E, 35, 93, 79, 39,   ,  5, 80, 91, 95, 39
```

このデータからデータフレームを作成すると、行の項目数が同じでないため、次のようなエラーが発生して、データフレームの作成が失敗します。

```
>k <- read.table("log.csv", header=FALSE, sep=",")
 以下にエラー scan(file, what, nmax, sep, dec, quote, skip, nlines, na.strings, :
   1 行目には 11 個の要素がありません
```

行の項目数が一定でないとき、足りない項目を欠損値で埋めてデータフレームを作成するには、次のように行います。この後の欠損値の扱いについては、前述の対応を行います。

```
>(k <- read.table("log.csv", header=FALSE, sep=",", fill=TRUE))
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12
1  A 85 82 44  9 74 52 64 45  11  90  NA
2  B 11 54 89 43 68 16 65 50  NANANA
3  C 50 71 93 16 34 62 NA  6  18  70  53
4  D 14 36 35 52 26 76  5 NANANANA
5  E 35 93 79 39 NA  5 80 91  95  39  NA
```

read.table 関数の fill 引数に TRUE を指定することで、列の項目の個数を揃え、足りない項目に欠損値を設定する。

【参考】セキュリティ対応に関して

処理対象のデータの中に、個人情報や機密情報の保護の観点で関わるものがあるとき、該当するデータの削除や他の値への変換を行う場合があります。値の変換に関しては、次の選択肢があります。

・匿名化　　個人やモノを特定する属性を処理し、特定ができないようにする
・仮名(かめい)化　　具体的な名称を記号に置き換える(オープン ID などの活用)
・暗号化　　情報の可読性を取り除く

このような処理は、R においては不得手なものであるため、その役割を外部プログラムに持たせることを考えます。

【参考2】Excel の「データ分析」アドオンを使用したデータ分析

データや分析手法によっては、別のツールを使って分析を行ったほうが早い、見やすい場合もあるため、場面によってツールを使い分けると、より効率的でわかりやすいデータ分析を行うことができるでしょう。

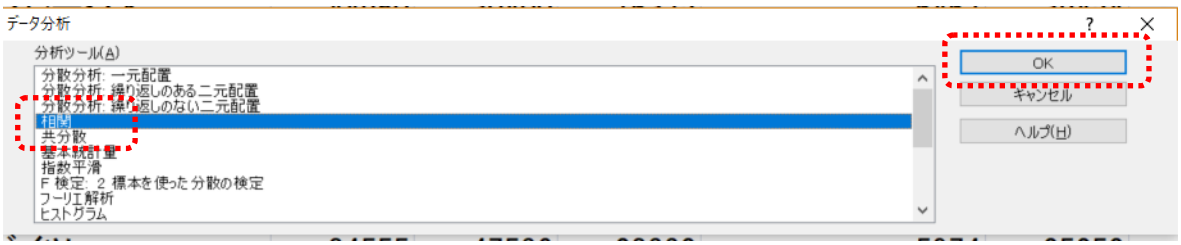
下記に、Excel の「データ分析」アドオンを使用したデータ分析例（ここでは相関、カラースケール、データの並び替え）を掲載します（Excel のバージョンによって多少メニュー配置などは変わります）。

1. 「2018_JP_CA_cnt_費目大区分ごとにまとめ.csv」を Excel で開きます。ここには観光庁が公開している、訪日外国人消費動向調査のデータを大項目ごとにまとめたデータが保存されています。

	A	B	C	D	E	F	
1	国名	宿泊費	飲食費	交通費	娯楽等サービス費	買物代	その
2	韓国	24974	19961	7635	3918	21549	
3	台湾	35312	28190	13548	5057	45441	
4	香港	45625	36887	16680	5064	50287	
5	中国	47854	39984	16835	7998	112104	
6	タイ	36836	27740	15033	4416	40248	
7	シンガポール	63311	41406	19890	6468	41691	
8	マレーシア	44950	30400	16372	6467	39424	
9	インドネシア	48117	29156	20945	5586	37598	
10	フィリピン	31448	30074	14459	6077	39595	
11	ベトナム	55818	43846	18901	5923	63650	
12	インド	75371	34026	21864	3746	26416	
13	英国	100692	56050	33172	8341	22641	

2. ツールバー内「データ」より「データ分析」ツールを選択します。

3. データ分析にて使用可能な分析ツールが表示されます。今回はこの中から「相関」を選択します。

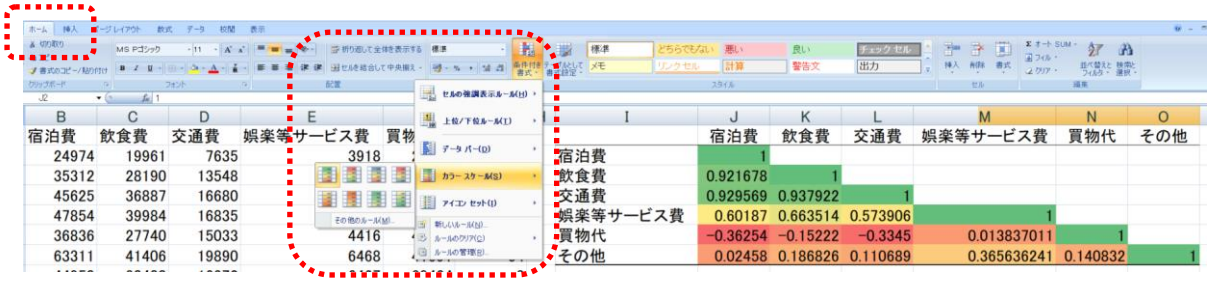


4. 入力範囲を選択します。右側の「範囲選択アイコン」をクリックし、B1 から G21 までを範囲選択し、「先頭行をラベルとして使用」チェックボックスにチェックを入れ、「出力先」を同じシート内の適当な場所に設定して「OK」を押します。

5. 総当たりでの相関分析結果が出力されました。

	宿泊費	飲食費	交通費	娯楽等サービス費	買物代	その他
宿泊費	1					
飲食費	0.921678	1				
交通費	0.929569	0.937922	1			
娯楽等サービス費	0.60187	0.663514	0.573906	1		
買物代	−0.36254	−0.15222	−0.3345	0.013837011	1	
その他	0.02458	0.186826	0.110689	0.365636241	0.140832	1

6. 「カラースケール」を使用して、より直感的に表示を行うことも可能です。相関係数の入っているセルをすべて選択し、ツールバーの「ホーム」から「条件付き書式」内「カラースケール」を選択し、わかりやすい色を適用します。

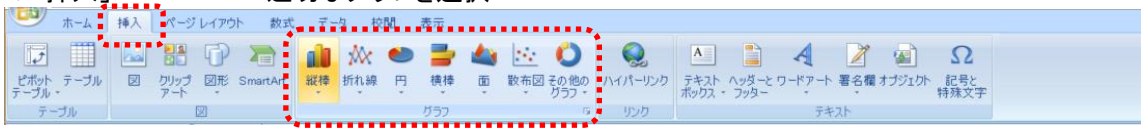


これにより、分析結果をより直感的に確認することができました。

また、データの並べ替え(「データ」メニューから「並べ替え」)や、グラフの挿入(「挿入」メニューから適切なグラフを選択)などにより、容易にデータの操作や可視化が行えるため、簡易な分析であればこれらを利用して分析結果を作成するのも良いでしょう。

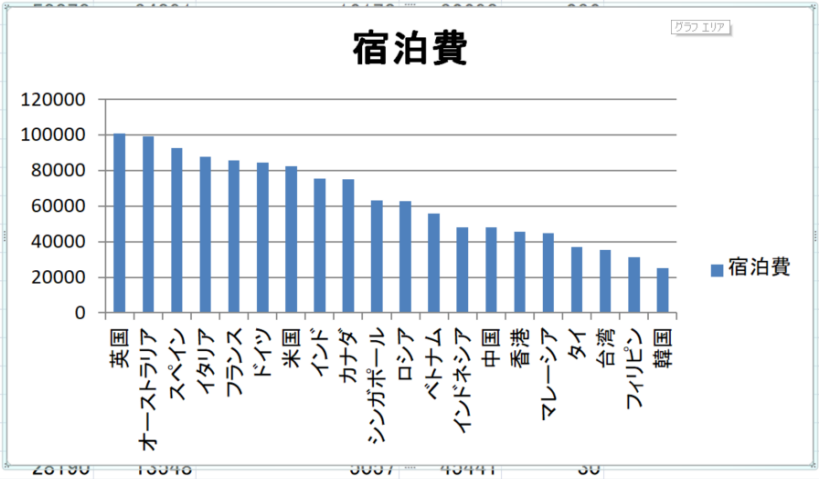
▼「データ」メニュー → 「並べ替え」

▼「挿入」メニュー → 適切なグラフを選択



下記は宿泊費を降順で並び替えし、宿泊費の列にカラースケールを適用、さらに国名と宿泊費で棒グラフを作成している例になります。

国名	宿泊費	飲食費	交通費	娯楽等サービス費	買物代	その他
英国	100692	56050	33172	8341	22641	34
オーストラリア	99175					
スペイン	92543					
イタリア	87652					
フランス	85544					
ドイツ	84555					
米国	82286					
インド	75371					
カナダ	74857					
シンガポール	63311					
ロシア	62710					
ベトナム	55818					
インドネシア	48117					
中国	47854					
香港	45625					
マレーシア	44950					
タイ	36836					
台湾	35312					
フィリピン	31448	30074	14459	6077	39595	268
韓国	24974	19961	7635	3918	21549	47



演習問題　解答例

第3章演習問題 –代表値-

(1)本の分類ごとの売上データを準備しました。各分類の代表値を求め、各分類の売上の特徴を分析してください。

【解答スクリプト例】

```
> # データのインポート
> x <- read.csv("prac3.csv")
> # データの概要確認
> head(x)
        日付 プログラム NW DB OS 統計
  1 2015/4/1          33 68 43 73    3
  2 2015/4/2          38 73 48 78    6
  3 2015/4/3          33 68 43 73    2
  4 2015/4/4          33 68 43 73    3
  5 2015/4/5          36 71 46 76    5
  6 2015/4/6          39 74 49 79    8
> dim(x)
  [1] 1461    6

> # 代表値の確認
> summary(x)
          日付         プログラム          NW              DB              OS              統計
  2015/10/1 :   1   Min.   :30.00   Min.   :65.00   Min.   :40.00   Min.   :70.00   Min.   :  0.00
  2015/10/10:   1   1st Qu.:33.00   1st Qu.:76.00   1st Qu.:42.00   1st Qu.:72.00   1st Qu.: 37.00
  2015/10/11:   1   Median :35.00   Median :80.00   Median :45.00   Median :75.00   Median : 53.00
  2015/10/12:   1   Mean   :36.27   Mean   :79.79   Mean   :44.88   Mean   :74.88   Mean   : 54.14
  2015/10/13:   1   3rd Qu.:39.00   3rd Qu.:84.00   3rd Qu.:47.00   3rd Qu.:77.00   3rd Qu.: 72.00
  2015/10/14:   1   Max.   :53.00   Max.   :98.00   Max.   :50.00   Max.   :80.00   Max.   :125.00
  (Other)   :1455
> range(x$プログラム);range(x$NW);range(x$DB);range(x$OS);range(x$統計)
  [1] 30 53
  [1] 65 98
  [1] 40 50
  [1] 70 80
  [1]   0 125
```

```
> colSums(x2<- x[,-1])
  プログラム          NW          DB          OS        統計
       52992      116571       65576      109406       79103

> sd(x$プログラム);sd(x$NW);sd(x$DB);sd(x$OS);sd(x$統計)
[1] 4.353605
[1] 5.919647
[1] 2.938578
[1] 2.938578
[1] 24.77376

> var(x$プログラム);var(x$NW);var(x$DB);var(x$OS);var(x$統計)
[1] 18.95388
[1] 35.04222
[1] 8.63524
[1] 8.63524
[1] 613.7391
```

●分類ごとの代表値

	プログラム	NW	DB	OS	統計
平均	36.27	79.79	44.88	74.88	54.14
中央値	35	80	45	75	53
合計	52992	116571	65576	109406	79103
最大	53	98	50	80	125
最小	30	65	40	70	0
件数	1461	1461	1461	1461	1461
分散	18.95388	35.04222	8.63524	8.63524	613.7391
標準偏差	4.353605	5.919647	2.938578	2.938578	24.77376

分類	考察
プログラム	平均が低いが、売上は比較的安定。安定しているため、継続的な売上は見込める半面、売上拡大は見込みづらい。
NW	平均が高く、売上は比較的安定。平均が高く安定しているため、主力商品となる。ただし、売上拡大は見込みづらい。
DB	平均がやや低く、売上は非常に安定。平均は低いが非常に安定しているため、確実な売上が見込める。ただし、売上拡大は見込みづらい。
OS	平均が高く、売上は非常に安定。平均が高く安定しているため、主力商品となる。ただし、売上拡大は見込みづらい。
統計	売上は平均的であるが、売上は非常にばらつきがある。ばらつきが多いため、様々な要因により売上が変化しやすいと思われる。売上が非常に高いこともあるため、売上拡大という視点で見た場合は、重点管理対象商品となる。

第4章演習問題 –相関分析-

(1)分類「統計」の売上増加の原因を調べるために、「ビッグデータ」のつぶやき数、分類「統計」の売上数のデータを収集しました。この 2 項目間に関係があるかを分析してください。

【解答スクリプト例】

```
> # データの読み込み、確認
> x <- read.csv("prac4.csv")

> # 必要部分のみを抽出、格納
> x2 <- x[,c(2,3)]

> # 相関係数の算出（列名を指定）
> cor(x2$つぶやき数,x2$売上,method="pearson")
[1] 0.6978791

> # 相関係数の算出（データフレームごと算出）
> cor(x2)
          つぶやき数      売上
つぶやき数  1.0000000 0.6978791
売上        0.6978791 1.0000000
```

●つぶやき数と分類「統計」の売上との関係に対する考察

考察
つぶやき数と売上は、「量的変数」*「量的変数」であるため、相関係数を算出する。 相関係数は、約 0.698 ということで、強い正の相関があることが分かった。 この結果を活かし、今後、以下の施策が考えられる。 ・つぶやき数から売上予測をし、予測値により目標設定や在庫管理に活用する。 ・つぶやき数を増やせば売上も向上する可能性があるため、つぶやき数を増やす方法を検討する。

【オプション】

(2)分類「統計」の売上が伸びている原因を分析するため、日付(Date)ごとの「ビッグデータに関連するつぶやき数(TW_Bigdata)」「ビッグデータ特集ページへのアクセス数(AC_Bigdata)」、「データベースに関連するつぶやき数(TW_DB)」「データベース特集ページへのアクセス数(AC_DB)」「アクセスした顧客の平均年齢(Age Sales)」を収集しました。それぞれの項目と分類「統計」の売上(Sales)との関係を分析してください。

【解答スクリプト例】

```
> # データの読み込み、確認
> x <- read.csv("prac4-2.csv")
> head(x)
      Date TW_Bigdata AC_Bigdata TW_DB AC_DB Age Sales
1 2011/4/1       1292       1540  1563  1613  48    70
2 2011/4/2       1320       1399  1260  1310  49    69
3 2011/4/3       1394       1281   893   943  39    74
4 2011/4/4       1399       1338  1043  1093  44    73
5 2011/4/5       1270       1396  1254  1304  41    69
6 2011/4/6       1292       1408  1301  1351  40    68

> # x2 に x の全行、1 列目以外を代入
> x2 <- x[,-1]

> # x2 の全変数間で相関係数を算出
> cor(x2)
            TW_Bigdata AC_Bigdata        TW_DB        AC_DB         Age        Sales
TW_Bigdata  1.00000000  0.4877878  0.018360919  0.018360919 -0.51923672  0.697879076
AC_Bigdata  0.48778784  1.0000000  0.733612228  0.733612228 -0.46375824  0.672506859
TW_DB       0.01836092  0.7336122  1.000000000  1.000000000  0.04187528 -0.009578731
AC_DB       0.01836092  0.7336122  1.000000000  1.000000000  0.04187528 -0.009578731
Age        -0.51923672 -0.4637582  0.041875275  0.041875275  1.00000000 -0.728051715
Sales       0.69787908  0.6725069 -0.009578731 -0.009578731 -0.72805171  1.000000000
```

●各項目と「統計入門」の購買（Sales）の関係に対する考察

項目	考察
TW_Bigdata つぶやき （ビッグデータ）	つぶやき(ビッグデータ)と売上の相関係数は、0.6979 と高い数値になった。つぶやき数(ビッグデータ)は、売上拡大や売上予測に利用できる要因といえる。
AC_Bigdata アクセス数 （ビッグデータ）	アクセス数(ビッグデータ)と売上の相関係数は、0.6725 と高い数値になった。アクセス数(ビッグデータ)は、売上拡大や売上予測に利用できる要因といえる。
TW_DB つぶやき （データベース）	つぶやき(データベース)と売上の相関係数は、-0.0096 と低い数値になった。つぶやき(データベース)は、分類「統計」の売上にはあまり関係の無い要因といえる。
AC_DB アクセス数 （データベース）	アクセス数(データベース)と売上の相関係数は、-0.0096 と低い数値になった。アクセス数(データベース)は、分類「統計」の売上にはあまり関係の無い要因といえる。
Age アクセス 平均年齢	アクセス平均年齢と売上の相関係数は、-0.728 と高い負の数値になった。負の相関であるが、アクセス平均年齢は、売上拡大や売上予測に利用できる要因といえる。

第5章演習問題 -回帰分析-

(1)「ビッグデータ」のつぶやき数と分類「統計」の売上の関係が強いということで、その関係を利用し、分類「統計」の売上予測をすることになりました。実績値をもとに回帰分析をおこない、分類「統計」の売上予測式を作成してください。

【解答スクリプト例】

```
> x <- read.csv("prac4.csv")
> head(x,3)
       日付 つぶやき数 売上
1 2011/4/1        1292   70
2 2011/4/2        1320   69
3 2011/4/3        1394   74

> res<-lm(売上~つぶやき数,data=x)     # 変数 res に分析結果を代入
> summary(res)                        # summary 関数で内容を確認
Call:
lm(formula = 売上 ~ つぶやき数, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-18.439  -6.320  -1.117   5.683  32.908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.413661   4.311551   1.488    0.138
つぶやき数  0.055419   0.002981  18.590   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.669 on 364 degrees of freedom
Multiple R-squared:  0.487,     Adjusted R-squared:  0.4856
F-statistic: 345.6 on 1 and 364 DF,  p-value: < 2.2e-16

> plot(x$つぶやき数,x$売上)
> abline(res,col="red")
```

●分類「統計」の売上予測式

売上予測式
売上数をy、つぶやき数をxとすると、予測式は以下になる。 **y = 0.0554x + 6.4137**

●回帰分析の考察

考察
予測式より、つぶやきが 1 つ増えるごとに、売上が約 0.0554 向上することが分かった。 これにより、つぶやき数の増加が売上向上の施策として考えられる。 しかし、決定係数は 0.487 と決して高いとは言えず、予測式の精度は高くないので、おおよその売上予測に使う程度にとどめておいた方がよい。

【オプション】

(2)分類「統計」の売上とそれらに関連があると思われる項目を利用し、分類「統計」の売上予測をすることになりました。実績値をもとに回帰分析をおこない、分類「統計」の売上予測式を作成してください。

【解答スクリプト例1】

```
> # データのインポート、確認
> x <- read.csv("prac4-2.csv")
> head(x,3)
        Date TW_Bigdata AC_Bigdata TW_DB AC_DB Age Sales
1 2011/4/1         1292         1540  1563  1613  48    70
2 2011/4/2         1320         1399  1260  1310  49    69
3 2011/4/3         1394         1281   893   943  39    74

> # x2 に x の全行、1 列目以外を代入
> x2 <- x[,-1]
> # x2 の全変数間で相関係数を算出
> cor(x2)
            TW_Bigdata AC_Bigdata        TW_DB        AC_DB         Age        Sales
TW_Bigdata  1.00000000  0.4877878  0.018360919  0.018360919 -0.51923672  0.697879076
AC_Bigdata  0.48778784  1.0000000  0.733612228  0.733612228 -0.46375824  0.672506859
TW_DB       0.01836092  0.7336122  1.000000000  1.000000000  0.04187528 -0.009578731
AC_DB       0.01836092  0.7336122  1.000000000  1.000000000  0.04187528 -0.009578731
Age        -0.51923672 -0.4637582  0.041875275  0.041875275  1.00000000 -0.728051715
Sales       0.69787908  0.6725069 -0.009578731 -0.009578731 -0.72805171  1.000000000

> # 関係性の強い変数を使ってモデルを作成
> res <- lm(Sales~TW_Bigdata+AC_Bigdata+Age,data=x2)
```

【解答スクリプト例2】

```
> # summary 関数で内容を確認
> summary(res)

Call:
lm(formula = Sales ~ TW_Bigdata + AC_Bigdata + Age, data = x2)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6641  -3.9663   0.0095   3.9341  22.2548

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.951497   6.363633   7.378 1.11e-12 ***
TW_Bigdata   0.026096   0.002617   9.971  < 2e-16 ***
AC_Bigdata   0.025654   0.002523  10.167  < 2e-16 ***
Age         -0.908429   0.072383 -12.550  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.148 on 362 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7413
F-statistic: 349.6 on 3 and 362 DF,  p-value: < 2.2e-16
```

●分類「統計」の売上予測式

売上予測式
売上数をp、つぶやき(ビッグデータ)をx、アクセス数(ビッグデータ)をy、アクセス平均年齢をzとすると予測式は、以下になる。 **p = 0.0261x + 0.0256y − 0.9z + 46.951**

●回帰分析の考察

考察
予測式により、「つぶやき(ビッグデータ)」1 増えるごとに 0.0261 売り上げが増え、「アクセス数(ビッグデータ)」1 増えるごとに売り上げが 0.0256 増え、アクセス平均年齢が1下がるごとに売り上げが 0.9 上がる傾向であることが分かった。 これをもとに、「つぶやき(ビッグデータ)」と「アクセス数(ビッグデータ)」の増加、アクセス平均年齢の引き下げが売上拡大の施策として考えられる。 決定係数は、0.7434 となり、予測式の精度は悪くないと言える。