Day4
データの要約と可視化

技術部データ基盤チーム 財津大夏 / GMO PEPABO inc.

2022.07.15 データエンジニアリング研修 基礎編 Day4

GMOペパボ

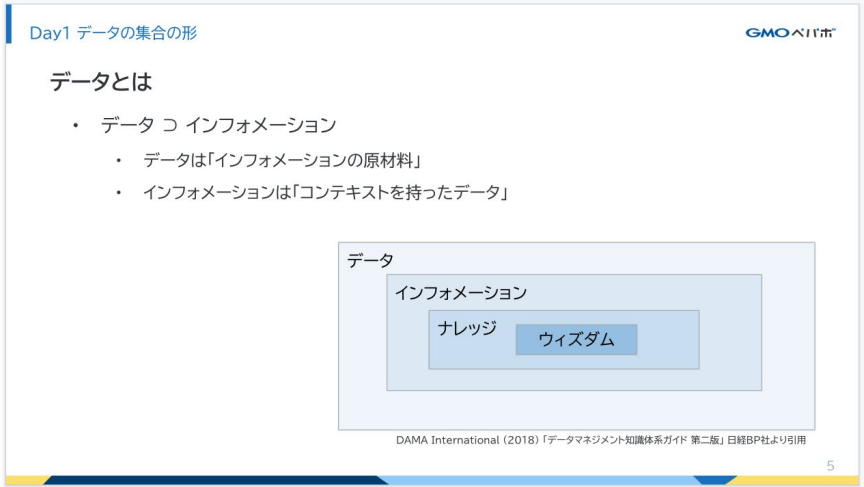
カリキュラム目標と概要

- **Day1: 扱いやすいデータの集合の形を理解できる**
  - データを構造化するための知識の導入
- **Day2: 初歩的な SQL を使ってデータベースからデータを参照できる**
  - データを参照するために必要な基礎的な知識の導入
- **Day3: 複数テーブルのデータを組み合わせて参照できる**
  - リレーショナルデータベースからデータを参照するための知識の導入
- **Day4: データを要約・可視化して情報や知識を取り出すことができる**
  - データを実際の施策や判断に利用するために必要な知識の導入

➡ 各日のハンズオンを通して手を動かしながら知識の解釈を高める

なぜ要約・可視化が必要なのか

- データは「インフォメーションの原材料」
- データを構造化することでインフォメーションに出来た
- インフォメーションを眺めてナレッジ（判断材料）を取り出すことは難しい
  - 10行くらいならまぁ…
  - 1万行のデータを目視する？
- 人間が理解できる形にしたい



データの要約

データを要約する

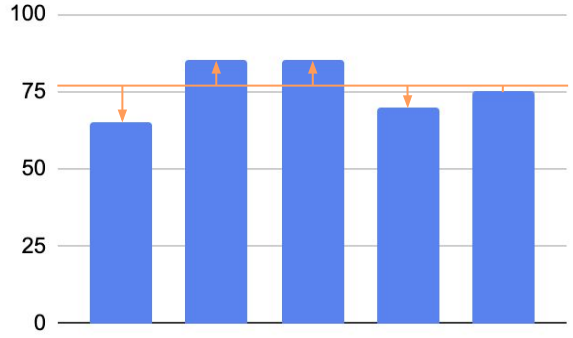
- データの集合の特徴を簡潔に表現する
  - 統計量による表現
    - 中心的傾向の測度: 平均, 中央値, 最頻値
    - ばらつきの測度: 分散, 標準偏差, 範囲, 四分位数, 四分位偏差
  - グラフによる表現
- データの集合同士の関連度合いを表現する
  - 相関係数

中心的傾向の測度

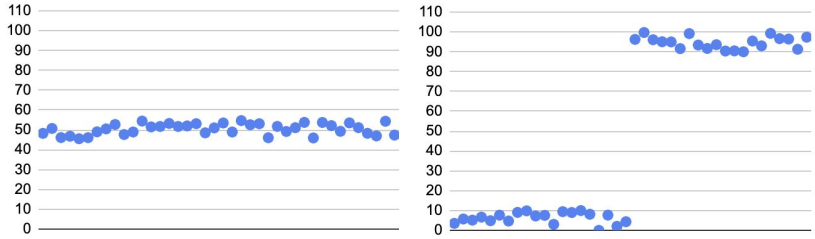
- 平均, 中央値, 最頻値
  - 例)あるクラスのテストの点数: 65, 85, 85, 70, 75
  - 平均(算術平均): 和を要素数で割った値
    - (65 + 85 + 85 + 70 + 75) / 5 = **76**
  - 中央値: データを値の大きさ順に並べたとき順位が中央の値
    - 65, 70, **75**, 85, 85
  - 最頻値: 最も頻繁に表れる値
    - 65, **85**, **85**, 70, 75

ばらつきの測度

- 分散, 範囲, 四分位数
  - 例)あるクラスのテストの点数: 65, 85, 85, 70, 75
  - 分散: それぞれの値と平均値の差の二乗の平均
    - $((76-65)^2+(76-85)^2+(76-85)^2+$
      $(76-70)^2+(76-75)^2)/5=\underline{64}$
  - 範囲: 最小の値と最大の値, その差
  - 四分位数: データを値の大きさ順に並べて4等分したときの区切りの値
    - 第1四分位数, 25%ile値
    - 第2四分位数, 50%ile値, 中央値
    - 第3四分位数, 75%ile値



統計量だけ見ていれば大丈夫？



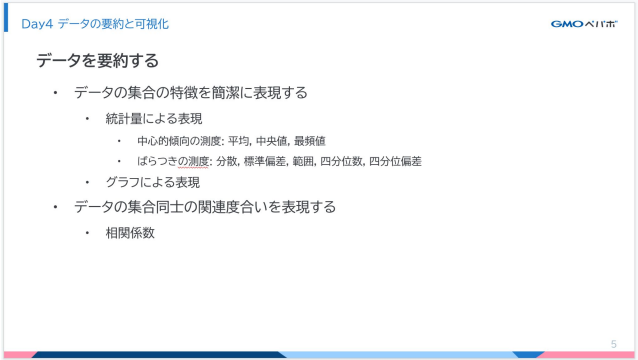
- 例)平均 50 の 2 つのデータセット
- まず分布を見ましょう
  - 分布が偏っていないか？
  - 外れ値は無いか？
- データセットのサイズを示しましょう
  - 対象件数が少なければ「偶然そうなっている」可能性も高まる
- 統計量の意味を理解して使いましょう
  - 「平均の平均」？

Standard SQL で統計量を出す

- 集計関数や集計分析関数で統計量を計算できる
  - 平均値: AVG()
  - 中央値: bqutil.fn.MEDIAN()
  - 最頻値: APPROX_TOP_COUNT()
  - 分散: VAR_SAMP()
  - 範囲: MIN(), MAX()
  - 四分位数: APPROX_QUANTILES()

記述統計学と推測統計学

- ここまでの話が記述統計学
  - 「観測されたデータはこうです」
- データから母集団を推測するのが推測統計学
  - 例）A/Bテストで観測された平均値に「差がある」
    - 偶然そうなったのではないか？
    - 「有意に差があるか」は計算によって明らかにできる



GMOペパボ

研修のあとに

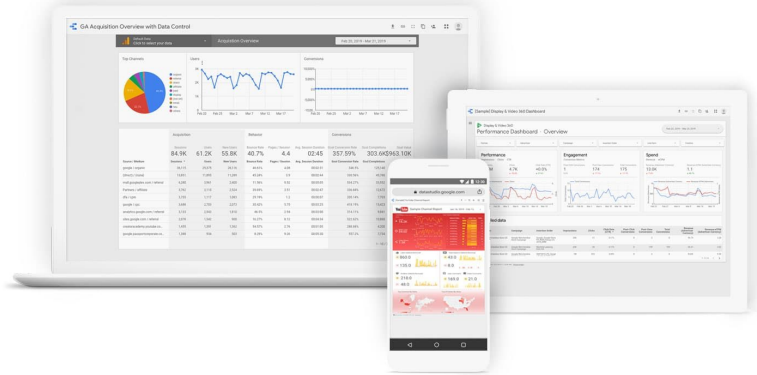
- データエンジニアリングのうちデータを操作・要約するための基本的な方法を学んだ
  - 観測されたデータを要約することができるようになった 💪
- データを操作して要約するだけでウィズダムに繋げることは難しい
  - 現実にデータを収集・分析・活用する上での考え方・技術が必要
  - 「データの集め方」「データの扱い方」「データの解釈の仕方」が分かると日々の仕事に自信をもってデータを役立てられます
    - 例）施策の効果をどう測定するべきか？[*1] 測定結果は偶然でないか？
    - 例）ユーザーアンケートをどう取るべきか？
  - データの収集・分析段階を網羅しているこの本が次の一歩にお薦め 👉
    - ほかにデータ基盤チームおすすめ書籍リストの 1-2 等級の書籍も

*1:「施策を実施した前後で比較」は比較していることにならないので、条件を統制して A/B テストを行いましょう。
機会損失を防ぎつつ良いパターンを探す多腕バンディットも各サービスに導入を進めています。

Google Data Studio による可視化

Google Data Studio*1

- グラフや表形式のダッシュボード, レポートを作成するツール
- BigQuery, Spreadsheets, Google Analytics などをデータソースにできる
- Spreadsheets や Slides と同様に Google Workspace で共有可能



画像は https://marketingplatform.google.com/intl/ja/about/data-studio/ から引用
*1: 日本では商標の関係で「Google データポータル」という名前になっています

おすすめのデータ処理の流れ

- Bigfoot（BigQuery）にデータを収集・集計
- Data Studio でダッシュボード, レポートにして可視化

➡ 収集・集計・可視化のフローを自動化できる

おすすめしないデータ処理の流れ

- Google Spreadsheets にデータを収集・集計・グラフ化
- グラフをキャプチャして Google Slides で作った会議資料に貼り付け

➡ 人手の作業が必要

使い方

- Google Cloud Self-Paced Labs をやってみましょう
  - データポータルを使ったデータ探索とレポート作成
    - https://www.cloudskillsboost.google/focuses/3614?locale=ja&parent=catalog

ハンズオン

ハンズオン

- 以下の内容を集計するクエリを作成してください。
  また、非集計の状態で結果を取得するクエリを作成した上で、
  結果を Data Studio に取り込み可視化してください。
  - ✍ (4-1) bigfoot org のリポジトリで Issue のラベルごとの件数
    - ヒント: labels は非正規化され、JSON 文字列として保存されています。
      JSON を加工するための関数が存在します
  - ✍ (4-2) suzuri と minne org のリポジトリで
    Workflow ごとに直近半年間の週別の Run の成功した回数と失敗した回数
  - ✍ (4-3) bigfoot org のリポジトリで Issue の作成時刻から
    初回コメント時刻までの時間の平均値と中央値

ハンズオン

- 以下の内容を集計するクエリを作成してください。

  また、非集計の状態で Data Studio に取り込み可視化してください

  - ✍ (4-4) bigfoot org のリポジトリの Pull Request の中で、

    renovate が作成したものについて以下の統計値:

    - 作成からマージされるまでの時間の平均値を分単位で(マージされたもののみでOK/以下同様)
    - 作成からマージされるまでの時間の中央値を分単位で
    - 作成からマージされるまでの時間の 90%ile 値を分単位で