令和8年度 市町村共通課題解決分科会（第2回） 開催レポート

- 日時：2026年9月29日（火）13:00〜15:30
- 会場：青森県庁 西棟 580 会議室
- テーマ：【整える】データクレンジング
- 登壇者：福田 藍至（ファシリテーター / 株式会社コンシス）

セッション項目	主な内容・ポイント
オープニング・前回のおさらい	・フォーム利用の意義：収集と集計が同時に完了し、転記ミスを防ぐ強み ・業務での活用：庁内業務（効率化）と庁外業務（市民の声収集）の活用事例 ・セキュリティ：ゼロトラストの考え方に基づく運用
データ可視化とBIツールの紹介	・**BI**ツールの役割：Looker Studio等を使用し、データをグラフィカルに「見える化」して高度な分析を行う ・悪いフォームの実例：自由記述や複数回答の多用がデータ汚染を招くことを実感
データの品質とデータクレンジング	・**Garbage In, Garbage Out**：ゴミデータからは無意味な分析結果しか生まれない ・品質の**6**要素：正確性、一貫性、適時性、有効性、一意性、完全性の解説 ・手作業の限界：置換機能などによる手作業での修正がいかに困難かを体験
Power Query・**AI**によるクレンジング実践	・**Power Query**：非破壊処理かつ直感的な操作でのデータ整形（条件列の追加等） ・生成**AI**の活用：自然言語による自動分類やエラー検出と「Human-in-the-loop」の重要性（ハルシネーション対策）
クレンジングの限界とフォーム設計の重要性	・完全性の限界：未回答などクレンジングでも補えない要素が存在する ・構造化データの取得：後からのクレンジング（地獄の作業）を避けるため、「粒度」を抑えた単一選択式の設問で集めることの重要性を強調

スライドと講義内容（録音文字起こし）

【オープニング・前回のおさらい】

【整える】
データクレンジング

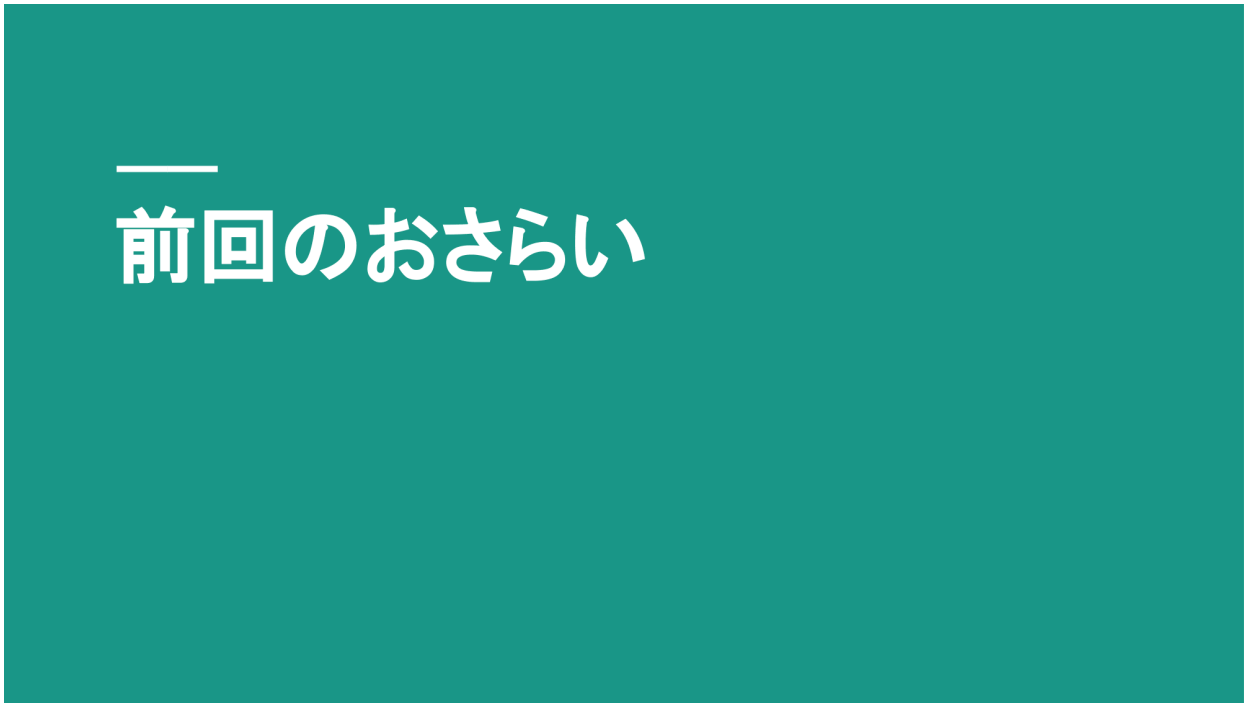
青森県庁市町村課｜市町村共通課題解決分科会・第2回

【福田】始めさせていただきます。第2回分科会で、今回のテーマは「整える データクレンジング」ということでやっていきたいと思います。よろしくお願いいたします。私、株式会社コンシスの福田と申します。今日のところはこのデータの整形の重要度を正しく認識するということです。

本日のゴール｜
データ整形の重要度を正しく認識する

まず、クレンジングという言葉を聞いたことがないという方、手を挙げていただいてもいいですか？は

い、あまりいらっしゃらないですかね。今回はあんまり面白くないです、地味です。ですが重要な話なので、ちょっと我慢してついてきていただければありがたいです。



前回のおさらいですね。全4回を通じてデータを集めてそれを活用することを皆さんと一緒にやっていきます。データを集めるとなると、今回はフォームを使って集めましょうということでやっていきます。

【おさらい】フォームによるデータ収集と利活用

なぜフォームなのか

- 「収集」と「集計」が同時に完了するから
- 回答者が送信した瞬間にデータベース化が完了
- 転記ミス・手作業ゼロ/ 予算ゼロ・手続きなしで今日から使える

こんな業務はもうやめましょう！

- 各課にWordやExcelの様式をメールで配って回収
- 届いたファイルを1冊ずつ開いて、手作業で1つのファイルに転記・集計
- 誤字脱字や入力漏れを1件ずつ目視でチェック、電話で確認する

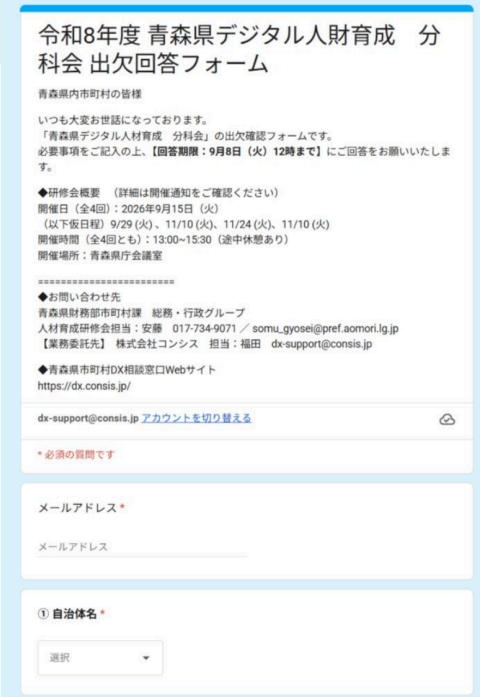
なぜフォームなのかということは、前回お話しした内容からすると、集めるのと計算するのが同時に完了するということですね。回答者がデータを送った瞬間にデータベースになっていくということで、人の手を介することがないので、基本的に転記のミスが生じないんですね。導入についても無料で使えるツールということで、便利なツールとして扱えるのでお勧めしておりました。そういった課題をフォーム一発で解決できますよというようなお話をしました。

【おさらい】
Googleフォームの使い方

集計用スプレッドシートと自動連携し、リアルタイムでのデータ可視化・分析が可能

AI・スプレッドシートとの強い親和性

回答が即座にデータベース化され、生成 AI（Gemini等）やBIツールと直接連動できるため、「データ収集から AI分析までの一連の体験」を学ぶのに適しています。



何を使うかというと、今回のフォームツールとしては Google フォームを使うということでお話ししてました。
いろんなフォームツールがあって、基本的な Excel みたいなスプレッドシートと自動で連携してリアルタイムで情報が集まっていく、それを分析できるという特徴があるんですけども、Google フォームについては Google の AI や Google スプレッドシートなど、グループウェアの中に全部入っているので、これらの製品を連携して使いやすいということで今回採用しておりました。

【おさらい】フォームの2大活用シーン

【庁内】手作業転記の「ゼロ化」

- 各課からの予算要望のとりまとめ
- 庁内研修・説明会の受講申込み
- 各種定期報告・アンケートの集約

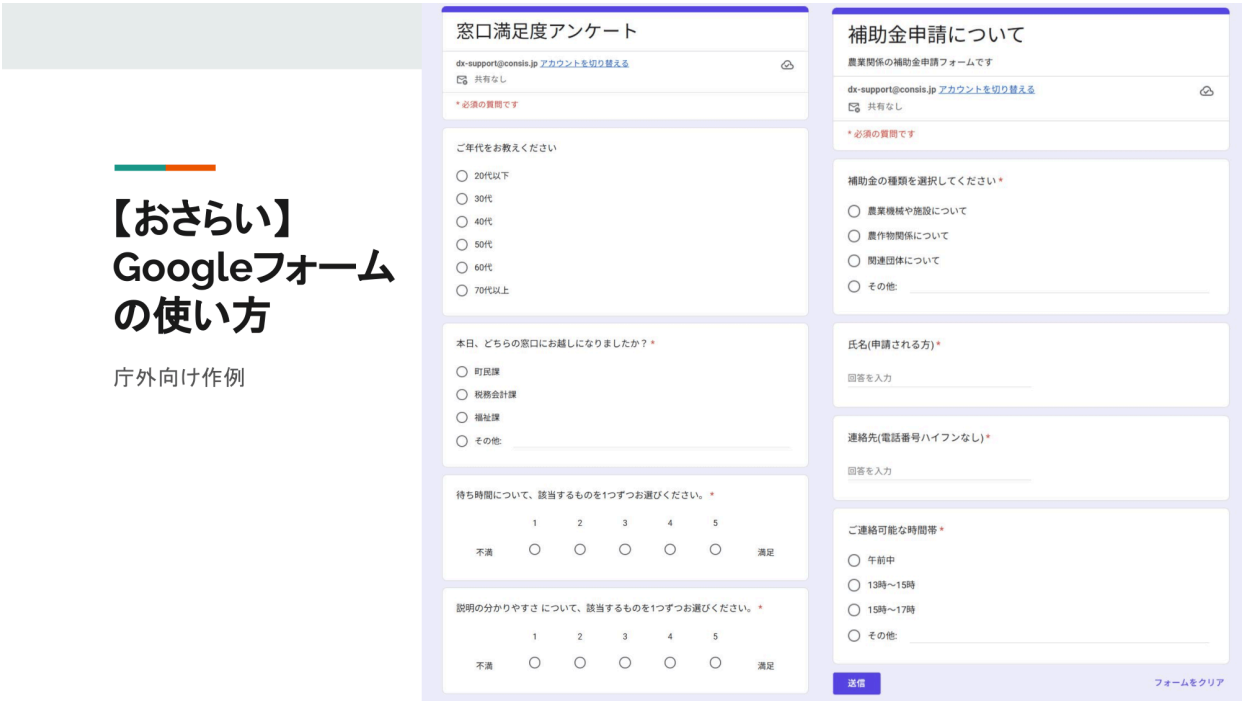
【庁外】市民の声を集める

- イベント参加者の満足度・属性調査
- 公園施設や道路の破損通報フォーム
- 地域の課題抽出・パブリックコメント

フォームを使うとなると、庁内の業務と庁外（市民対象）の内部・外部の業務に大きく2つ分かれるかなというところですね。
庁外は分かりやすいですよね。イベントのアンケートとか市民の声を集めるという性格のものですね。従来、紙でやってたようなアンケート代わりにこのフォームに入力してくださいね、というようなやり方で全然いいと思うんですけども。
庁内の業務に使うというところで言うと、先ほどお話ししたような決算等による転記が発生する作業とか、これを置き換えることができます。



皆さんに実際作っていただいた作例として2つ紹介したいと思います。
1つは庁内向けの業務効率化ということで、管理者にPCの管理をお願いする依頼です。前はどうやってたんですかね、メールとか紙でやってたんですかね。仮にそうであるなら、このフォームがあることによって、そういう面倒な作業が一網打尽にできますよね。同じように時間外勤務の申請で、これも送ったその場でドンと誰々が何時何分にというタイムスタンプがつくので、管理がしやすくなる一例だと思います。



こちら（もう1つ）は市民とか取引業者向けに作った感じのフォームですね。これについては窓口満足度アンケートとかはだいぶ分かりやすいですよね。多分現行でも同じような形で運用されてると思います。補助金の申請の基本関係等書いてますけども、結構電話で問い合わせを受けたりすることがあるかもしれないですが、フォームで入力してくださいねとすると、電話を受けてそれを手打ちして管理するという2プロセスから3プロセスぐらい省ける想定ができますね。皆さんだいぶ良い感じだと思います。

【おさらい】グループワーク＆ハンズオン

作成されたフォーム

業務効率化（77%）：通勤届、検査依頼、PC権限申請、時間外申請など

市民の声集約（23%）：敬老会アンケート、窓口満足度など

ご意見・ご感想

「DXには関わりがないと思っていたが、思いのほか簡単に活用できると感じた」

「スプレッドシートが思っていたよりも色々なことができそう」

「初めて使ったが使いやすかった。業務の効率化につながると思う」

今、庁内の業務で苦労してること、面倒だと思ってることとか、こういうことをフォームに使えばいいんじゃないかなということをグループでお話ししていただいて。簡単でいいんですよ。いきなり難しいこと始めなくていいので、まず簡単ですぐできることから始める。それが実感いただけたのはすごい良かったかなと思います。スプレッドシートについても、連携するやり方をするとだいぶいろんなことができるので、そこにも気づきを感じていただいたのは良かったかなと思います。

【おさらい】グループワーク＆ハンズオン

主なハードル：

- 時間・リソース:
日常業務が忙しい、テスト運用に不安(6名)
- スキル不安: Googleフォームの細かい操作方法(6名)
→できる範囲でOKです。
「知る・ふれる・試してみる」をできたのが重要！
- セキュリティ: 個人情報や部外秘データの取扱(5名)
→ゼロトラスト（許可した人だけがアクセスできる）を徹底！

懸念と課題：

- 「セキュリティ的（LGWAN環境など）に庁内で使うことが難しそう。」
- 「庁内の規則等の問題で、実際にデータ収集に使用することはできない。」

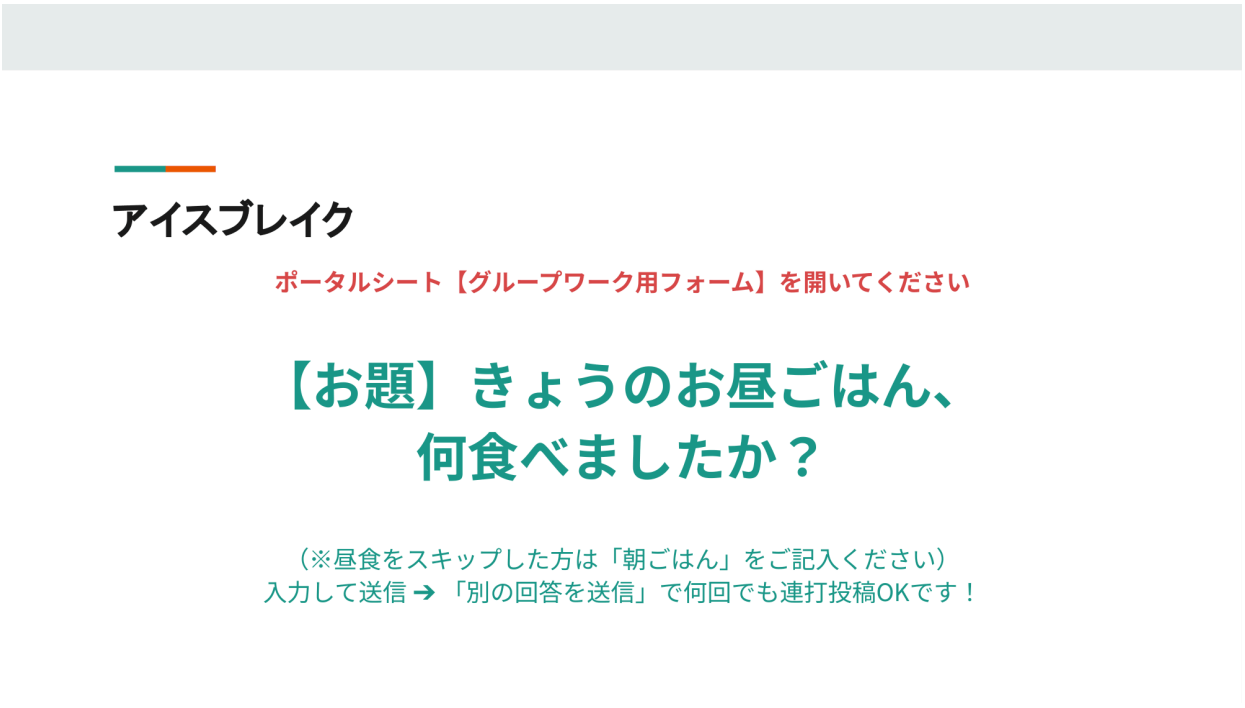
→LoGoフォームなど使用可能なツールで試してみましょう！

逆に「これはちょっと…」というお声もありましたね。一番はやっぱり日常の業務が忙しいということと、

早速試運転で走らせるのに若干不安があるという方。あとはフォームを作るのに細かい操作がちょっと不安だという方が多かったと思うんですけど、これもできる範囲で本当に大丈夫なので、まずやってみたということが重要かなと思います。やってみて「ここは使えるかな」「ここはもうちょっと足りないかな」みたいに、とりあえず手触りだけ実感いただければ十分かなと思います。

やっぱり自治体の仕事だとセキュリティというところが一番ネックになると思うんですけども、基本的に Google のサービスというのは「ゼロトラスト」という考え方をしてるんですね。誰もアクセスできないようにする、許可・アクセス権を付与した人にだけ通すよという設定で、これを設定で解除はできるんですけども、基本的にゼロトラストで運用するということであれば、割合安心して使えるかなと思います。
今回の事前に共有したポータルシートも同じようにゼロトラストの運用で、前回参加された方はそのまま行けたと思うんですけども、端末を変えられたり、今回初めて参加された方は、まず「アクセス権のリクエストをしてくださいね」と出ましたよね。認めた状況で初めてアクセスができるというゼロトラストというのも、ちょっと覚えていただければいいかなと思います。

実際に動かすとなると、LGWAN環境とかがネックになる規則もありますよね、ルール化されてないから使えないというところで。今回 Google フォームでやりましょうというお話だったんですけど、もし実際の業務に使えないとなると、使えるもの、例えば LoGoフォームとかを使えると思うので、やり方を覚えた上で使えるツールに置き換えて試していただければのかなと思います。



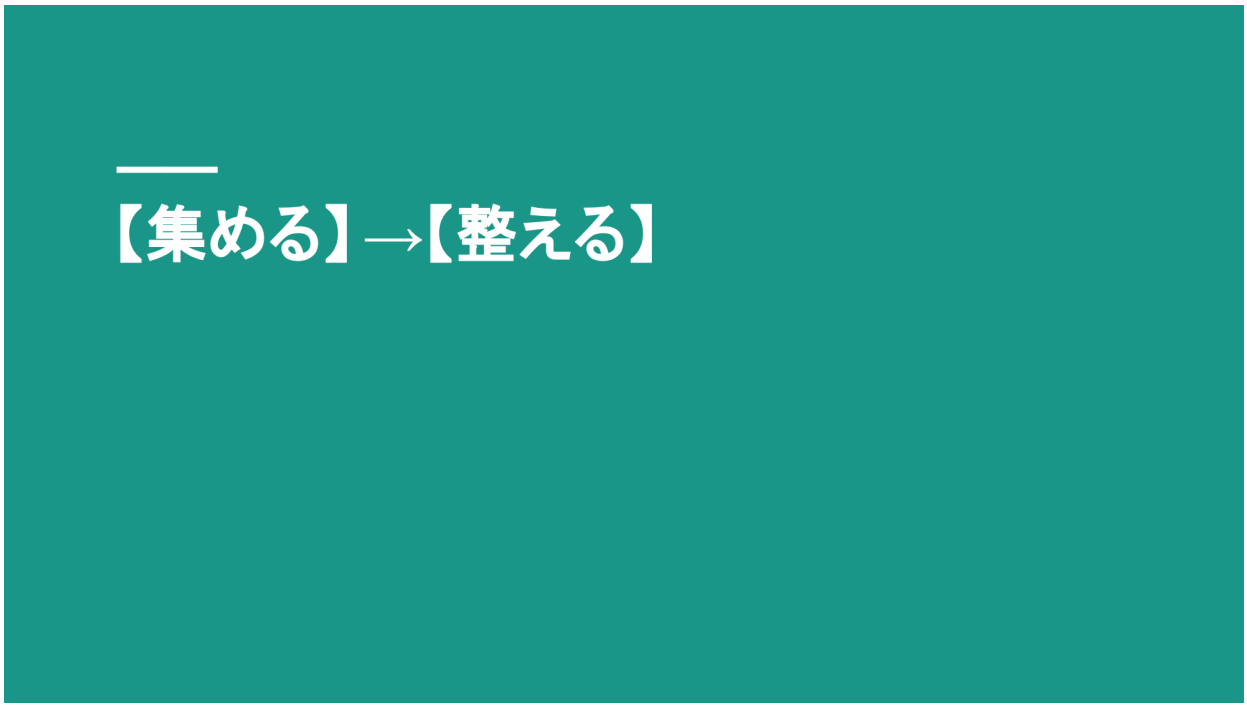
前回もやったんですけども、アンケートを取りたいと思います。ポータルシートのリンクがありますよね？1番上のリンクをちょっと踏んでいただけますか？フォームが出てくると思うので。



A	B
Form_Responses	
タイムスタンプ	おひるごはんアンケート
2026/09/29 13:01:49	コンビニのサンドイッチ
2026/09/29 13:22:15	コンビニおにぎり
2026/09/29 13:22:17	牛丼と豚汁ラーメンセット
2026/09/29 13:22:19	担々麺
2026/09/29 13:22:19	ラーメン
2026/09/29 13:22:23	カレーパン
2026/09/29 13:22:23	中華そば
2026/09/29 13:22:24	ナポリタン
2026/09/29 13:22:24	ツナカレーホットサンド
2026/09/29 13:22:25	ZOOTの塩豚骨ラーメン

今日の昼ご飯、何食べましたか？お、早いですね。サンドイッチ、おにぎり、おにぎり。前回はラーメンが多かったですよね。この「牛丼豚汁ラーメンセット」って熱いですね。すごい眠くなりそうですけど。12…26と、あと1件2件の回答を見たらちょっと移っていきますかね。ホットサンド、カレー味のツナなんですかね？これもコンビニですか？ZOOTの塩豚骨、気になりますね。はい、ご協力ありがとうございました。

【データ可視化とBIツールの紹介（Looker Studio / Power BI）】



【福田】じゃあちょっと今日のお話をしていきます。前回はデータを集めるというお話でしたけど、今回はそれを「整える」ということですね。
この分科会を開くにあたってちょっとオーダーがあって、BIツールを紹介してほしいというようなことが

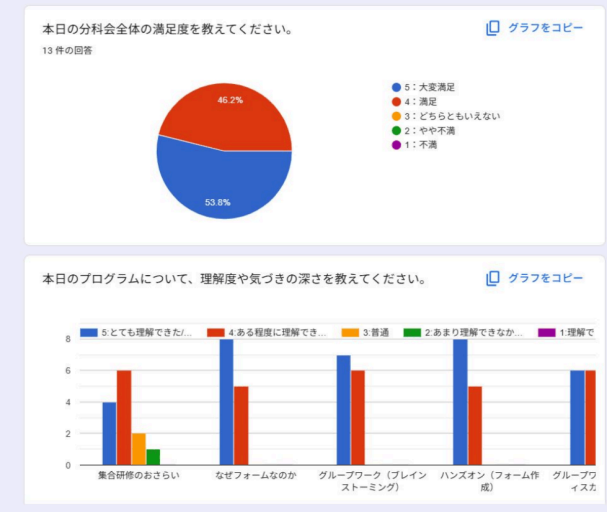
あったので、このお話をしたいと思います。

BI（Business Intelligence）ツール

収集したデータを可視化・視覚化

例）googleフォームの回答要約機能

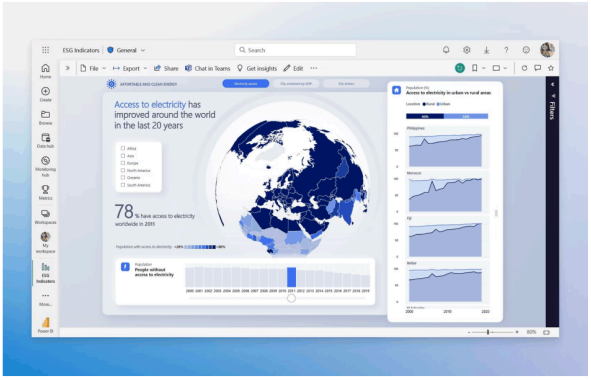
※基本的な集計データを確認するには必要十分



BIをご存知ですか？ご存知の方いらっしゃいますか？BI、ビジネスインテリジェンスツールと言うんですけど、集めたデータを「見える化」するツールですね。見たことある人いると思うんですけど、前回受講された方は絶対覚えてると思うんですが、Google フォームの「回答の要約」タブがありますよね。これを開くと回答がグラフで自動的に出てくるということで、これも BI ですね。

基本的な、例えば前回の分科会のアンケートなんですけども、「満足度を教えてください」で「4」が多くて、概ね良かったんだなというのは見れば一発で分かるので、これでも一つの手段なんですけども、高度な分析が必要になってくるとちょっとこれだと足りない。例えばクロス分析だったり、複数のデータを掛け合わせてどういう変化をしてくだろうというところまで見ようとなると、BIツールを使う必要があります。

BI（Business Intelligence）ツール



収集したデータを可視化・視覚化

クロス分析や複数データの掛け合わせなど、
精緻な分析が可能

主なBIツール

- Microsoft Power BI
- Googleデータポータル（旧Looker Studio）

など

有名どころだとMicrosoftのPower BIとか、Google のデータポータルですね。前はデータスタジオ、その前がデータスタジオで名前がコロコロ変わって紛らわしいんですけど、今回は一番新しい名前でご紹介します。外部のデータを引っ張ってきて、それをグラフィカルに表示できるというものですね、この仕組みは共通です。

今回使うBIツール

- （比較的）簡単に作成可能
- 無料で利用できる
- 共有・連携機能



cf) MS Power BI

- 関数の知識がやや必要
- 無料で利用可、共有・連携は有償版のみ

今回は Google のデータポータルを使っていきたいと思います。理由としては、『比較的』簡単に作

成可能であること。Microsoft の Power BI 等と比べると割合簡単にできるということで、ただやっぱりちょっと専門の知識も必要になってくるので括弧付きです。それと「無料で使える」「共有と連携の機能が充実している」。この3点ですね。
Microsoftはいけないことはないんですけども、データポータルよりも関数等の知識がちょっと必要かなというのと、無料で使えるんですけども共有や連携は有料版のみという条件がありました。

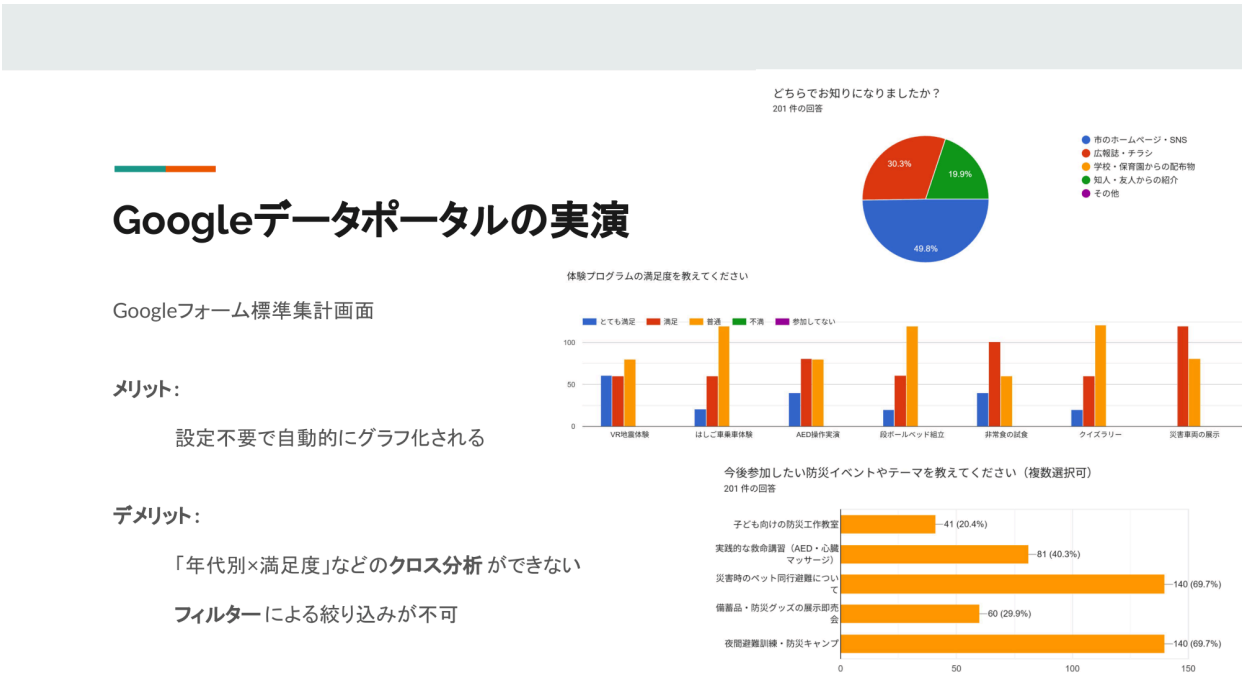
（Power BIのサンプル資料を表示）
これが Microsoft Power BI で作った資料なんですけども、色々高度な分析ができるんですけど何が致命的かと言うと、これ僕、無料のアカウントで使ってるんですけど、今こうやって皆さん見れますよね。これは僕のパソコンがここにあってプロジェクターに出してるから皆さん見れてるんですね。リンクを発行して「このリンクから見れますよ」というのをやるには有料版を使わないといけないんですよね。これが結構ダメかなと思います。

試しに皆さん検索窓で「データポータル」と検索してみてください。素っ気ない画面が出てくると思うんですけども、「使ってみる」に進んでみてもらえれば。Google アカウントの認証を要求されると思うんですけども、Google アカウントに入っているデータを引っ張ってくるから事前にOKをくださいねという認証ですね。最悪開けなければ今は全然大丈夫です。画面に出てきて「作成」というところですね。真っ白い画面が出てきて、「何のデータを取り込みますか？」というのを求められるわけですね。めちゃめちゃあります、1400個とかいっぱいあるんですけど、大体使うのってこの辺り（Googleスプレッドシートなど）ですね。スプレッドシートを開くと、今認証してる Google アカウントに入っているスプレッドシートが出てきます。まだ進めなくて大丈夫です、こんな感じだよというのが分ければ今日は全然大丈夫です。こんなものがあるんだなというぐらいで大丈夫です。



ちょっと詳しく実演したいと思います。シートを開いていただいて、架空のイベント「〇〇市ファミリー

防災フェスタ」のアンケートがあるんですけど、実際に回答してみてください。適当で大丈夫ですので。
展開する側、管理画面も見れるので管理画面も見てみてください。



管理画面も回答が進んだ方、見てみてください。選んでみると全部で211件の回答がありますねというのが分かりますね。ちょうどいいグラフで何が何なのかも見られるようになってます。
これだけでも十分いいかなと思うんですけども、一番いいのは何もしないで勝手にこういうグラフを作ってくれるということが一番のメリットですよね。

なんですけども、ここからさらに踏み込んだ、例えば「年代別の満足度は違うのか」とか「20代と40代を比べて回答の内容が違うのか」とか、そこまではこのグラフだけだとちょっと見えないというところで、もしそういう高度な分析が必要であればデータポータルを使ってみましょうということですね。
シートにあるデータポータルのリンクを開いてみてください。

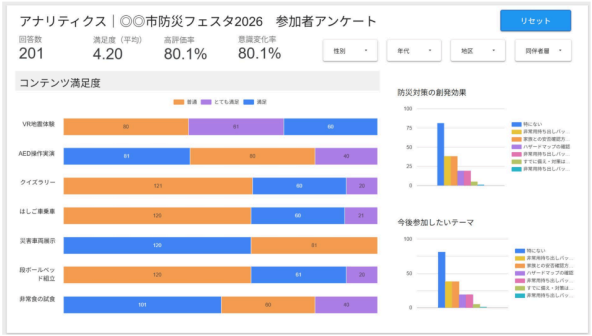
https://datastudio.google.com/reporting/d59f0b77-edd8-4551-937a-806281c0e7f9

Googleデータポータルの実演

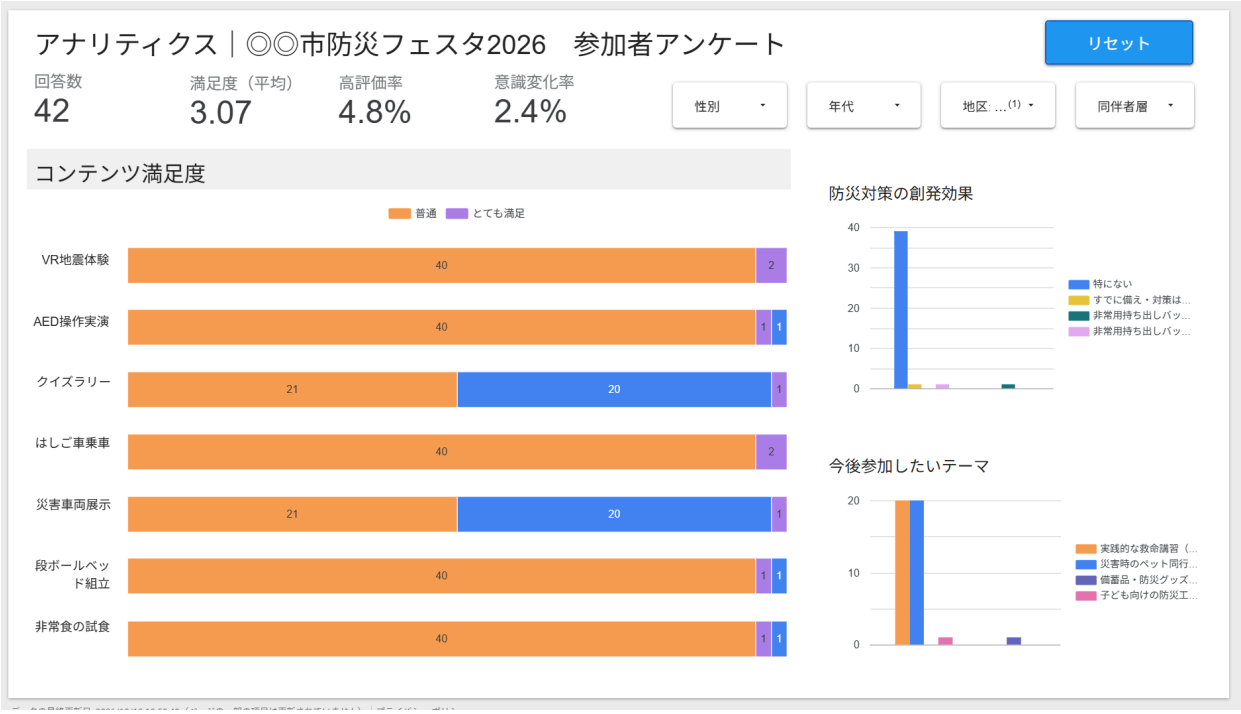
BIの強み

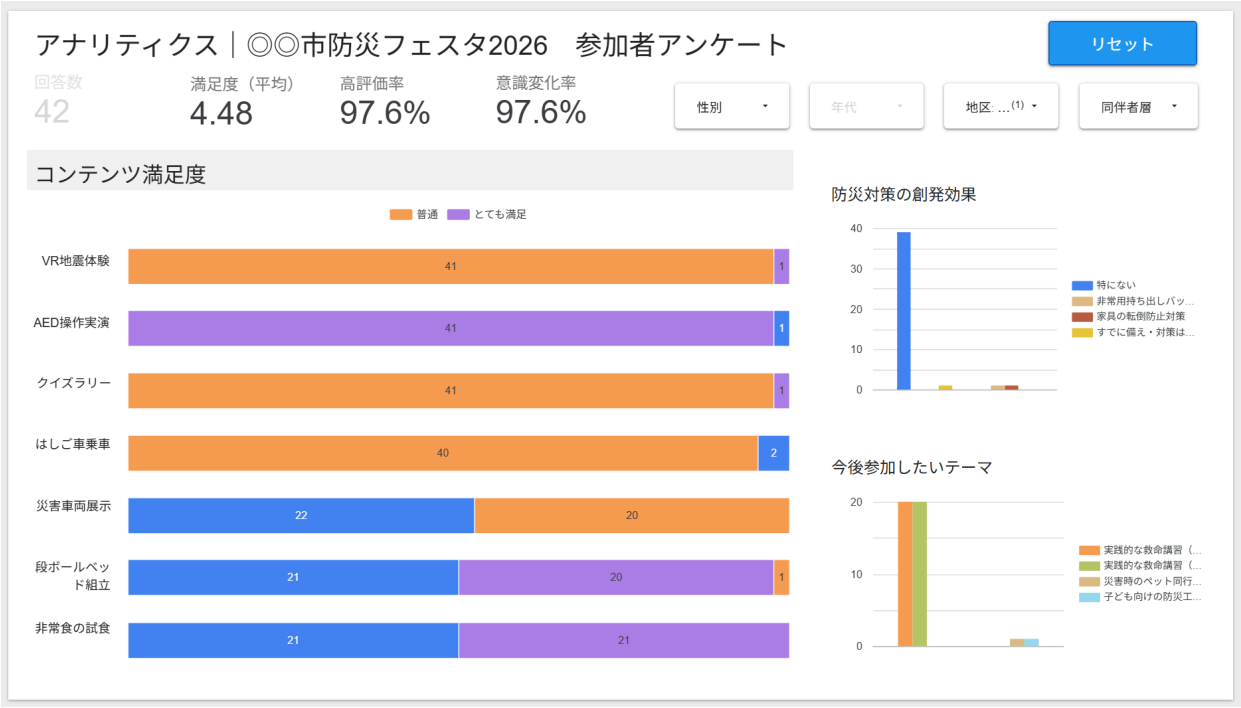
- フィルター機能
- リアルタイム計測

高度な分析が可能に



何ができるかと言うと、コンテンツ満足度とかがありますね。例えば地区を絞ってみましょう。ここに「高評価率」とか「意識変化率」とありますよね。意識変化率というのは「このイベントを通じて防災意識が変わりましたか？」に対して「変わった」「まあまあ変わった」という人を抽出してパーセンテージにしてるんですけど、この「中央地区」の人たちはそのあたりが異様に低いなというのが分かりますよね。





に対して「北部地区」の人は意識変化率が100%なんですよね（※画像は事後、分科会参加者の回答を含んだため97.6%になっています）。高評価率も高い。平面の回答分析だと分からないんだけども、BIを通すと「中央地区」の人だけ防災に対する意識がめちゃめちゃ低いということが分かったりしますよね。これがフィルター機能ですね。

性別（男女でどう違うのか）、年代（20代と40代と60代以上でどう違うのか）、地区でどう違うのかというフィルターをかけて回答を分けて見ることができるということが強みです。

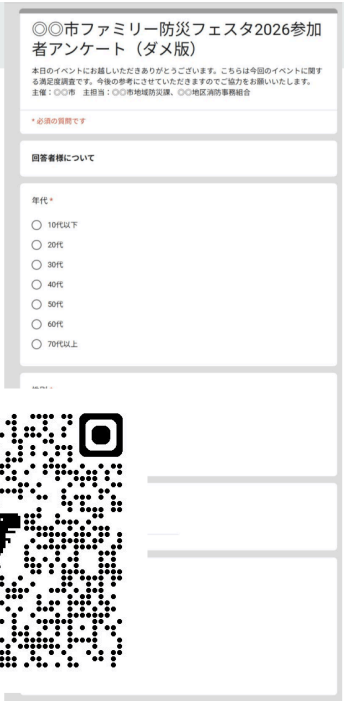
ここに回答数が「201」と出てるんですけども、今見ると「202」になってますよね。完全なリアルタイムではないんですけども、元のデータが更新されるのに連動してここも変わるんです。ちゃんと更新が反映されていく。なので高度な分析がこれでできるようになるということですね。

もうひとつのアンケート

https://docs.google.com/forms/d/e/1FAIpQLSfruUNLMi_kirWDOjG0WXyrGWAR9nhjeQKGL9e0u18alqIujg/viewform?usp=header

先のフォームと、どこが違う？

実際に回答して、見つけてみてください。



ここでもう1つのアンケートがあります。リンクを開いてみてください。
内容はだいたい同じなんですけども、なんか違うんですよね。ここに「ダメ版」って書いてるぐらいなんで良くないフォームなんですけども、前回皆さんにフォーム作っていただく時に「回答者が答えやすい形を意識して作ってくださいね」というお知らせをしたと思うんですけども、その観点も投入して何が違うかをちょっと見つけてみてください。

このフォーム、どこがダメですか？

- 「お住まい」が**自由記述**
- 「体験プログラムの満足度」が**複数選択可＆無回答可**
- 「始めたい対策」が**自由記述**
- 「参加したいテーマ」が**自由記述**

→データの正確性・完全性が確保できない！



【福田】何がダメだったか、皆さんなんとなくお分かりだと思うんですけども、この4つですね。

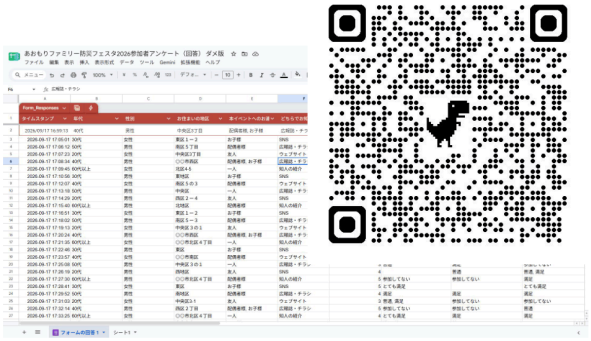
1. お住まい（住んでいる地区）：選ぶ形じゃない（自由記述になっている）。ダメ版じゃない方は選択肢5つぐらいで選べるようになってましたよね。
2. 体験プログラムの満足度：複数回答できちゃう。「とても満足」と「参加していない」を同時に選択できたり、スキップもできる。
3. イベントを通じて始めたいと思った対策
4. 今後どういうイベントに参加したいか
   この後者の2つも自由記述になっている。

設問が迷走していて、データの正確性と完全性がこのフォームだと担保できないですね。この正確性と完全性というのはあとから説明します。

見比べる｜収集されたデータ



https://docs.google.com/spreadsheets/d/1tGDK6xJbkSJ18CmrY0ZMEXxvec3O5iMnUBJYssXXWjE/edit?usp=sharing

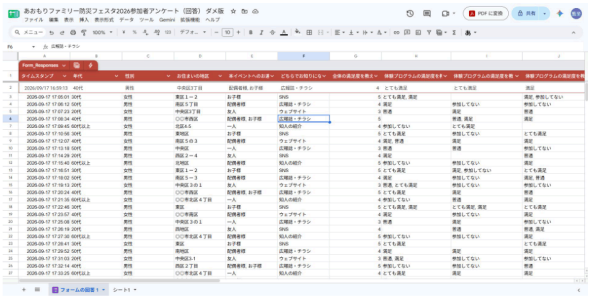


https://docs.google.com/spreadsheets/d/1cVG_OnAU6emp2gCEx645HyrPVopbCeOonC8B7ktwzO4/edit?usp=sharing

実際に集まったデータがどう違うかというのを見比べてみてほしいんですけども、シートの中にあるリンクで集計先のスプレッドシートを開いて見てみてください。
良い方が紫色のバーのシートで、ダメな方がお茶みたいな色してる方ですね。

- ◎◎市防災フェスタ2026参加者アンケート（回答）
- ◎◎市防災フェスタ2026参加者アンケート（回答）ダメ版

見比べる｜収集されたデータ



お住まい：自由記述
→**表記ゆれ**

PG満足度：複数選択可＆無回答可
→**データなし＆無効回答が混入**

始めたい対策：自由記述
参加したいテーマ：自由記述
→**回答が個別化して集約不可**

＝汚れたデータ（分析の信頼性が低い）

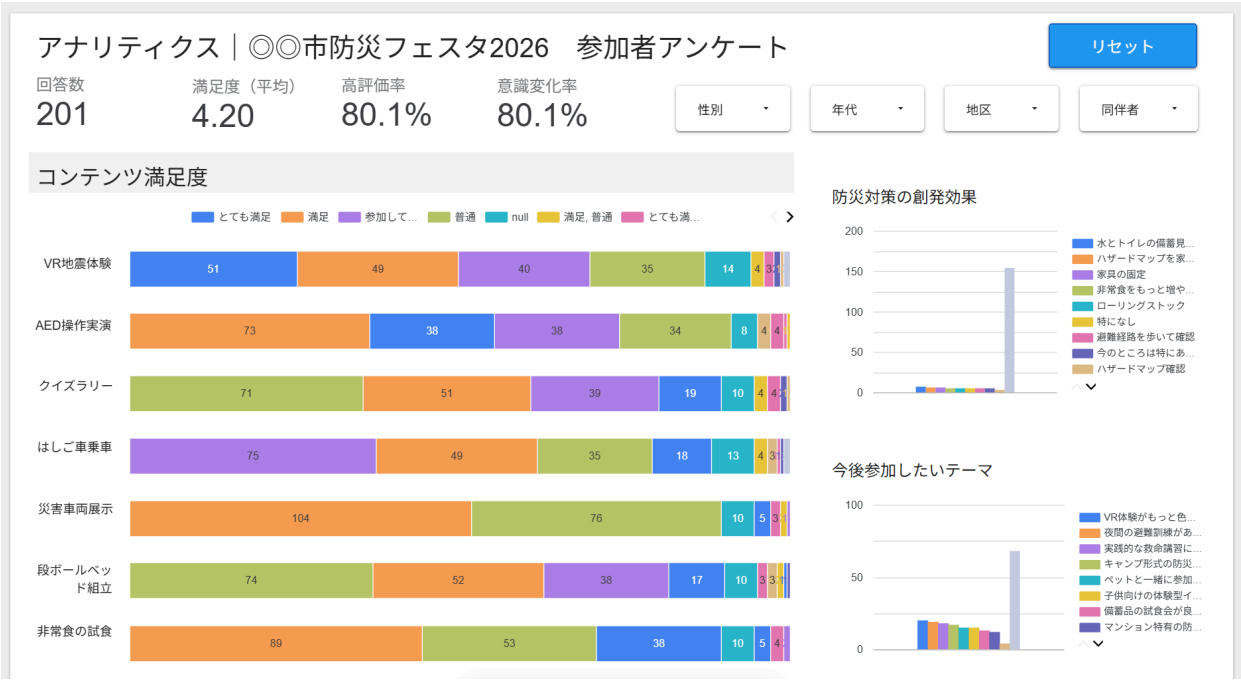
見ていただいて、さっきの4つのダメなところの項目に着目いただきたいんですけど、結構すごいことになってますよね。お住まいとかがめちゃめちゃですよね。プログラムの満足度も歯抜けになってたり、2つ回答して「満足」と「参加していない」が混ざったり、わけ分からないデータになってしまっている。
始めたい対策・参加したいテーマも、このいまだとそれぞれ個別の回答という扱いになっちゃって、まとめることができないですね。

ダメ版で集計した、これがまさに汚染された汚いデータということですね。これだとろくな分析ができないということです。

【データの品質（6つの基本要素）とデータクレンジング】



【福田】こういうデータの集計の仕方をしないためにはどうするかというところなんですけども、できちゃったらもう仕方ないので、これをどうにかしないといけない。まず一回BIに通してみましょう。
https://datastudio.google.com/reporting/bd227631-fe78-4fa0-b2be-aae811b37df7



一見して満足度がバグってるのはお分かりいただけましたか？こことかやばいですよね。なんでこうなってるかと言うと、複数回答をそれぞれカウントしちゃってるんですよね。本来であれば段階評価で

きることがこうなっちゃってるので何が何だか分からないと。地区もすごいですね、もうやばいことになってますね。何も分からない。
参加したいテーマとか「どういう気づきがありましたか？」というのも、全部個別回答になってて集計できないから全部「その他」になってしまっている。
「ハザードマップを家族で確認」と「ハザードマップの確認」、これ同じことですよね。なんですけども、個別の回答という認識をされちゃってるので、これだとせっかくBI使っても正しく情報が視覚化できてない、意味がないということですね。



これはITの世界でも警句みたいなものなんですけど、「Garbage In, Garbage Out」。文字通りですね。さっきゴミみたいなデータを入れたらゴミみたいなBIしか出てこない。これだと意思決定ができないので、データの品質で意思決定の精度が変わってくる。後に見たBIで何が分かるかと言うと、ほぼ何も分からないですね。なので、ちゃんとデータを整えることが必要になっていく。ここはなんとなく皆さんお分かりいただけたんじゃないでしょうか。

データの正確性と完全性が担保されないというお話をしましたけど、正確性・完全性とは何なのかというところです。データの品質を見定める上で、この6つの基本要素があるんですね。

データ品質｜6つの基本

① 正確性	データが事実に則しているか	❌ 顧客の引越し前の古い住所がそのまま残っている	⭕ 実際の住所とシステム上の住所が完全に一致している
② 完全性	必要なデータが欠けずに揃っているか	❌ 必須入力であるはずの「メールアドレス」が空欄（NULL）になっている	⭕ 分析に必要な項目がすべて埋まっている
③ 一貫性（整合性）	複数のシステム間でも矛盾がないか	❌ 社内システムAとシステムBで、同じ顧客の電話番号が違っている	⭕ どのシステムを見ても同じデータが同期されている
④ 適時性（最新性）	必要な最新データが利用できるか	❌ 昨日の売上データを見たいのに、反映されるのが3日後になる	⭕ 意思決定をしたいタイミングで、常に最新の鮮度のデータがある
⑤ 有効性（妥当性）	決めたルールや形式に合格しているか	❌ 年齢の欄に「-5歳」や「1000歳」といった値が入っている	⭕ 定められた桁数、日付フォーマット、選択肢（マスタ）の範囲内である
⑥ 一意性	データに重複がないか	❌ 同じ顧客が「会員ID：001」と「002」で二重登録されている	⭕ 1つの実体（人やモノ）に対して、データも1件だけ存在する

難しく書いてますけど、そんな難しくないです。

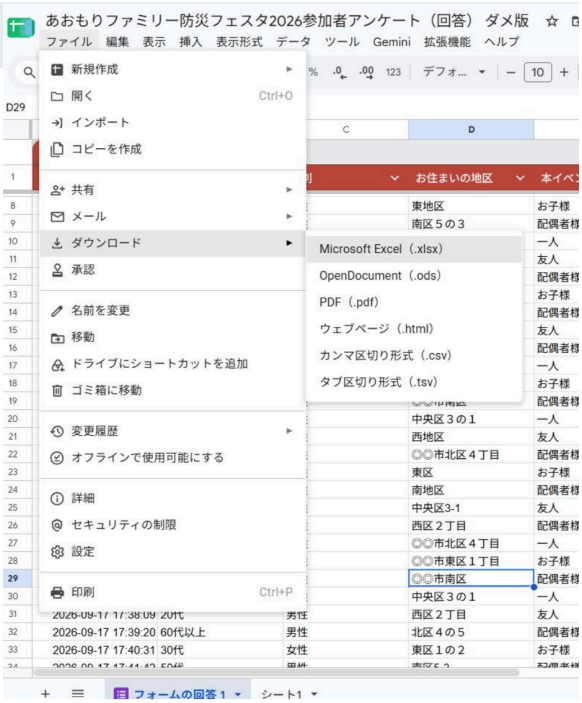
1. 正確性：Aというデータなのに**B**と書いてあったりする、事実を映していないということ。
2. 一貫性（整合性）：データを同期するにあたって、本来同じものなんだけどもシステムによって番号が違ったりしている。
3. 適時性：フォームで集まったデータはタイムスタンプが押されるので基本的に悩むことはないですが、タイムスタンプがないとしたら「いつのデータなのか分からない」。最新のデータと取得した当時の内容が変わっている可能性がある。いつのデータか分からないと正確性にも影響してくる。
4. 有効性：決められたルール。例えば年齢のところで「10代」「40代」「50代」というところにマイナスはないですし、「千歳」と入れたりするのもない。それが入っちゃっている状態。
5. 唯一性（一意性）：データに重複がないか。例えば皆さんがネットで買い物をした時に、今日頼んだものと昨日頼んだものは発送の注文番号は違うけれど、頼んだ「自分」というデータは同一ですよね。これが昨日頼んだ注文と今日頼んだ注文でそれぞれ別の会員として扱われるとなると、唯一性がおかしいということになります。
6. 完全性：データが欠落していないか。

この6つが欠けなく揃ったデータを目指しましょうということですね。この6つのどれかが足りない場合は、そこは頑張って補って綺麗にしましょうというのがデータクレンジングです。

データクレンジングの実践

汚れたデータをローカル環境で洗浄（整理）してみましょう

.xlsx形式でDLしてMicrosoft Excelで開きます



ということで、皆さんにデータクレンジングをちょっと実践していただきたいと思うんですけども、さっきのダメ版のスプレッドシートを開いていただいて、ブラウザ上でもいいですが、今回 Excel に落としたいので「ファイル」→「ダウンロード」から Excel 形式でダウンロードしてみてください。
お手持ちの端末に Excel が入っていないという方いらっしゃいますか？なければそのまま「コピーを作成」でダメ版のコピーを作って、そっちで操作してみてください。

誤った表記を正しいものにします。手作業でやろうとするとどうなるか：

- 手打ち修正：手打ちでセルをクリックして間違ったところを手打ちで直す。これはとりあえずないなと思いますよね。できるかもしれないけどやっていられないですよね。
- コピー＆ペースト：正しい値をコピーして間違ったところに正しく貼り直す。手打ちよりはいいけど、これもやっていられないですよね。

データクレンジングの実践

手法1：置換（Ctrl＋H） 誤った表記を正しい表記に置き換える

検索する文字列	置換後の文字列	
◎◎市	（空白）	「◎◎市」を除外
中央	中央地区	表記の統一
中央区	中央地区	**※注意：** 中央地区 が 中央地区地区 にならないよう置換順序に注意
せいぶ	西部地区	ひらがな表記の修正
東区	東部地区	表記の統一
（全角スペース）	*(空欄)*	末尾や間に挟まった見えないゴミスペースの削除

- 置換処理：**Ctrl + H**：検索する文字列を入れて置き換えたい文字列を入れる。

例えば「〇〇市」というのを全部まとめて消したいとなると、置換後の文字列は空白のままエンターすると「〇〇市」が全部消えますよね。同じく、中央地区を指している表記の中でも「中央」とか「中央区」とかが混ざったりしていますよね。これを全部「中央地区」に置き換えたい。「中央」を検索して「中央地区」に置換するかとなると、1回やっていただくと分かると思うんですけど、「中央地区区」みたいになったりするんですよね。面倒くさいなと。「中央区」を検索する文字列に入れて「中央地区」に置き換えるか、あるいはひらがなが混ざっていたら漢字にしてあげようか、変なスペースが入っていたら空欄に置き換えてゴミ（スペースやピリオド等）を削除してあげようかと。

データクレンジングの実践

手法1：置換（Ctrl＋H） 誤った表記を正しい表記に置き換える

検索する文字列	置換後の文字列
非常用持ち出し袋	非常用持ち出しバッグの準備
非常用バッグと家具固定	非常用持ち出しバッグの準備 家具の転倒防止対策
家具固定	家具の転倒防止対策
水と食料	水・食料の備蓄（3日分以上）
安否確認方法を決める	家族との安否確認方法の取り決め
特にない	*（空欄または「なし」）*

自由記述の方も表記がブレているので、全部「非常時持ち出しバッグの準備」という言葉に置き換えたいなと。2つ答えているものがあればコンマで区切るようにして「この2つの回答だよ」と置き換えをしますか。「水と食料」とかも正規の表現をしましょうか…

データクレンジングの実践

手法1：置換（Ctrl＋H） 誤った表記を正しい表記に置き換える

ちょっと無理 …

- 作業量が多い
- 完全な自由記述には置換ルールを 1件ずつ考えて設定
- 単純な置き換えだけでは正しきれないデータが多い

→ 大量かつ複雑なデータに対しては現実的ではない手法

…と、なるんですけど、やってみて無理じゃないですか？「できる」と思った方いらっしゃいますか？できるんであれば全然いいと思うんですけど、良くはないですね。地獄の作業なんですよ、こういうことは。

自由記述も置き換えるとなった時に、いろんな置き換え候補があるんですけど、まずこれをいちいち考えるのがだるすぎますよね。あと、このやり方だと一発で置き換えることができないものもありますし、単純に置換するだけだと拾いきれないデータが多すぎるので、200何件あるデータに対しては現実的ではない手法ですね。
あえてやっていただいたんですけど、実際にご自分で作業されると「これはダメだな」と実感できると思います。

【Power Query（パワークエリ）を活用したクレンジング実践】

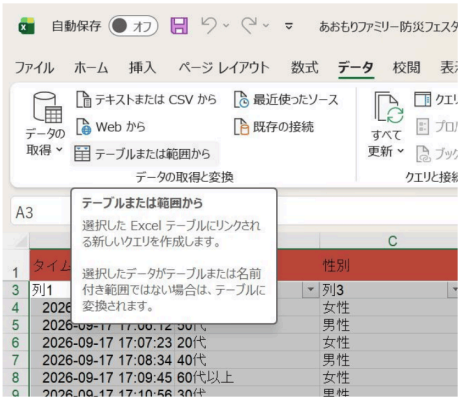


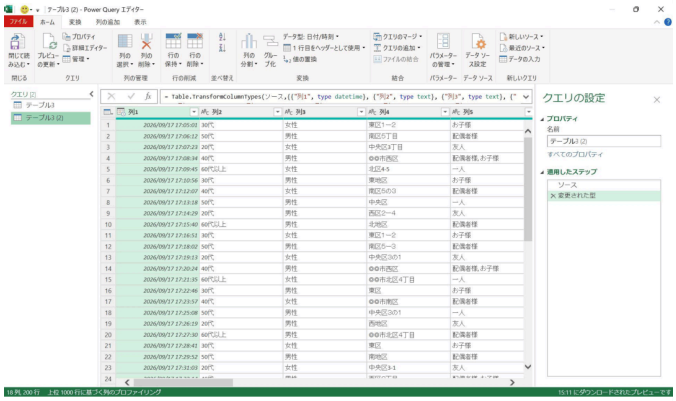
【福田】手法2として Power Query（パワークエリ）を使って整形する方法を紹介したいと思います。Power Query を知っている方いらっしゃいますか？これは単独のアプリではなく、Excel や Power BI に組み込まれているツールなんですけど、外部のデータに接続してそこで変換したり組み合わせたりができるんですね。

接続するのがポイントで、元のデータはいじらないんですよ。あくまで繋いでそこでやるので、元のデータはそのままの状態で保持されます。なので非破壊処理ができるということですね。元データに変更を加えることなく処理ができることと、関数の知識がなくてもある程度できること。

さっき「置換で直そうとなった時に関数使えばいいじゃん」と思った方いらっしゃると思うんですけど、関数が分からないとその処理ができないですよね。ですが Power Query を使うと、関数の知識がなくても関数の処理のようなことが直感的にできる、これが強みですね。

データクレンジングの実践｜ Power Query

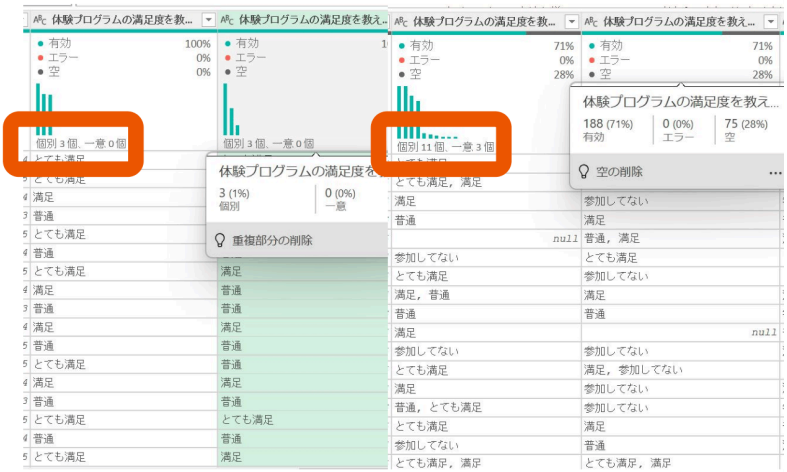




使い方ですが、Excel を開いていただいて、「データ」タブのリボンから「データの取得」または「テーブルまたは範囲から」どちらでもいいんですけど。「データの取得」→「ファイルから」→「Excel ブックから」で、さっきのファイルを開きます。インポートを押すとこういう画面が出てきます。
ここに出ているのは読み込んだ Excel ファイルの中身のプレビューです。全部ではないですが、大体こんなデータですよというのが出ます。ここから「データの変換」を押します。これが Power Query エディターですね。ここまで行ってみてください。

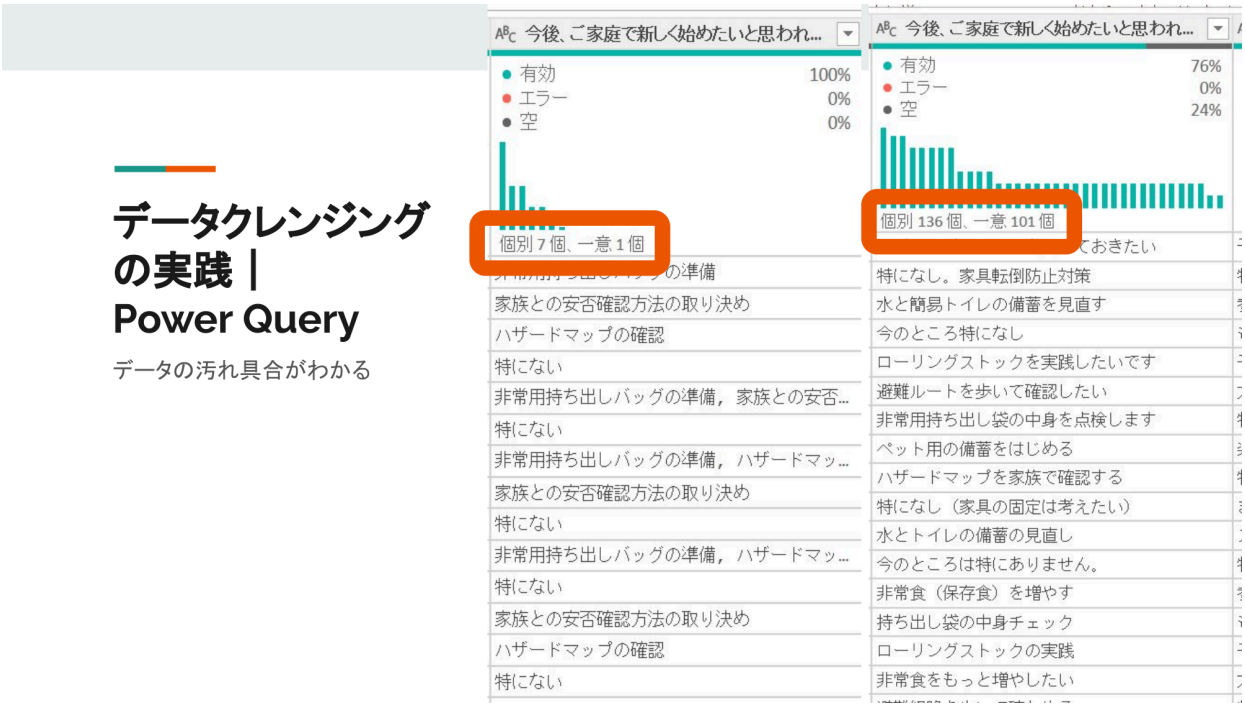
データクレンジングの実践｜ Power Query

データの汚れ具合がわかる



画面を見ると、列ごとに上にグラフが見える感じですか？出てない方は「表示」タブの「列の品質」「列

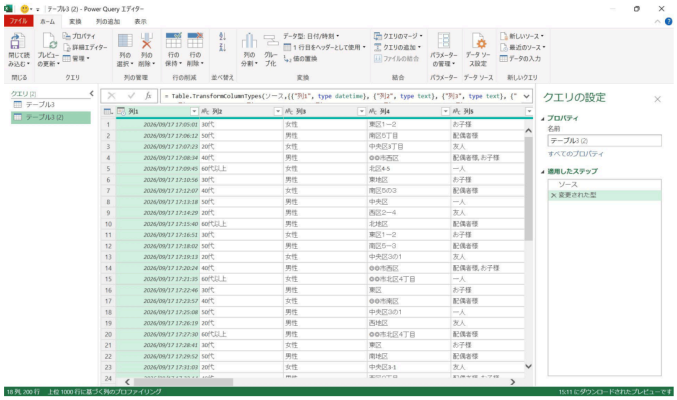
の分布」にチェックを入れると出てきます。
これ何かと言うと、文字通り品質なんですけど、例えばお住まいの地区というところ、綺麗な方のシートだと5分類でしたよね。なんですけどダメ版だと46通りもあると。満足度に関しても本来5段階評価なんだけど11の選択肢があると。



こちらは何通りでしたっけ、6か7ぐらいの選択肢だったのが136個もあるということで、これを見るとデータのヤバさ、綺麗さが一発で分かります。ただし良い方のシートは適切な数に収束しているということですね。正しい方だと7通りなんですけど、やばい方が136項目ということで、データクレンジングというのはこの136を正しく7個にまとめる作業と捉えていただけると間違いないかなと思います。

データクレンジングの実践｜Power Query

実演・条件列



どうやって揃えるのか実演していきたいと思います。例えば住所で「中央3丁目」から「中央」を取りたいんですよね、これは中央地区だし。「〇〇市」とかが入ってるやつもありますよね。これ置換だとどうにもならないんですけど、Power Query を使って「条件列」というのをやっていきます。

列名が「お住まいの地区」ですね。演算子が色々あるんですけど、「値を含む」を選びます。値に「中央」と入れます。出力する値を「中央地区」とします。



これで1回やってみます。
すると、ここに出てきたのが「中央」と入っているものが全部「中央地区」に置き換わりました。
「null」というのは回答がない、存在しないデータになっているということです。

0個	個別4個、一意0個	個別46個、一意2個		
	男性	中央区3丁目	中央地区	西
	女性	東区１－２	null	お
	男性	南区５丁目	null	西
	女性	中央区3丁目	中央地区	方
	男性	◎◎市西区	null	西
	女性	北区4-5	null	－
	男性	東地区	null	お
	女性	南区５の３	null	西
	男性	中央区	中央地区	－
	男性	西区２－４	null	方
	男性	北地区	null	西
	女性	東区１－２	null	お
	男性	南区５－３	null	西
	女性	中央区３の１	中央地区	方
	男性	◎◎市西区	null	西
	女性	◎◎市北区４丁目	null	－
	男性	東区	null	お

この要領でまた入れていきたいんですけど、「適用したステップ」にさっき条件列を追加して新しい地区の列を作った履歴が残っています。歯車マークを押すと、指定した列の条件をさらに追加できます。さっきダメだった「東」「南」「西」「北」が選ばれていなかったので追加していきましょう。



動かせる人はご自分の端末でも同じ操作をしてみてください。「クエリの設定」のところは最初つきづらいですが、触っているとなんとなく分かってくると思います。
開けない方は申し訳ないですがスクリーンをご覧ください。
仮に「中央東〇丁目」という住所があったとして、「中央」も「東」もつくけれど実は「東地区」なんだよねという時は、順番を上に入れ替えてあげてください。「上へ移動」。
この順番にすると、まず「東」が入っているものは全部「東地区」になります。なので「中央東」もまず「東地区」に変わる。で「中央」が残りますよね。東の処理が終わったから次に「中央」を見て「中央地区」に変わる。
なので「中央東」の東地区は置き換えられて、残った「中央」が中央地区に置き換えられてという風にある程度コントロールできますね。

個別 4 個、一意 0 個	個別 46 個、一意 2 個	
男性	中央区3丁目	中央地区
女性	東区１－２	東部地区
男性	南区５丁目	市外
女性	中央区3丁目	中央地区
男性	◎◎市西区	西部地区
女性	北区4-5	北部地区
男性	東地区	東部地区
女性	南区５の３	市外
男性	中央区	中央地区
男性	西区２－４	西部地区
男性	北地区	北部地区
女性	東区１－２	東部地区
男性	南区５－３	市外

一通り打ち終わってOKをすると、こんな感じでクレンジングした内容に置き換わります。大体綺麗になってると思います。できる方はこのあたりまで1回やってみてください。

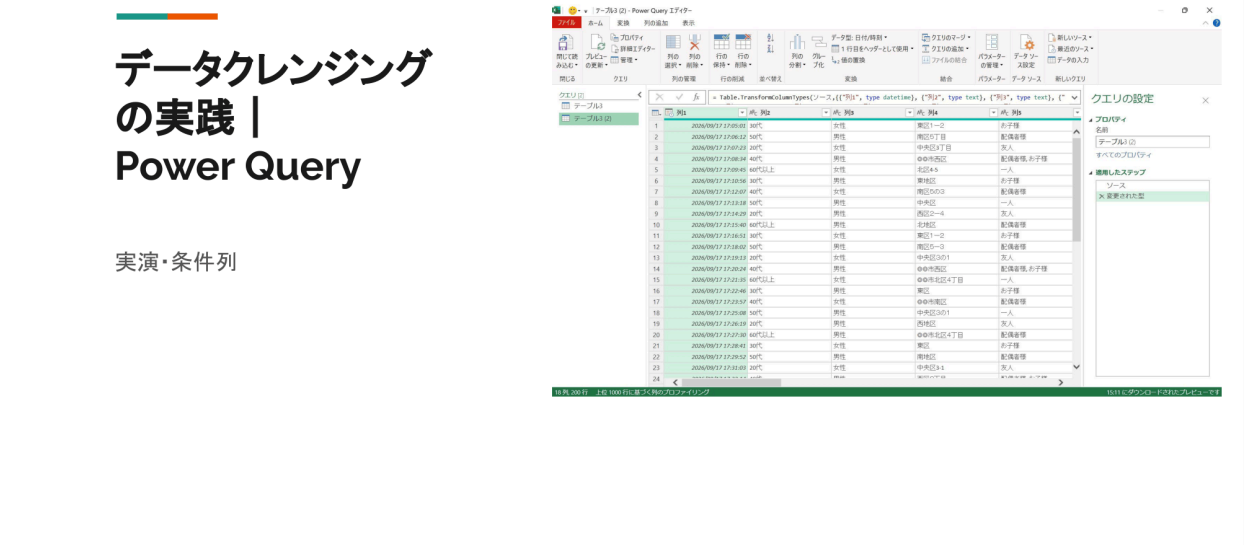
ここで10分ほどお休みを取りたいと思います。ご自身で条件列の操作をやってみたい方は続けてやってみてください。35分からまた再開したいと思います。ご質問がある方はご質問いただければと思います。

【休憩時間の質疑応答】

【受講者A】「表示」から触ってるんですけど、これが出なくて…設定が違うんですかね？
【福田】表示のところの変換ですね。条件で文が入ってないやつはこうする、というの指定してあって。実は関数で書くとこういう形なんですけど、それを代わりにポチポチでやってくれてるという感じですね。
【受講者B】「列の品質」というボタンがあるんですけど、無くてですね…まだ Excel の画面のままでした。
【福田】あ、そうだったんですね。「データ」から開いていただいて…この画面ですね。この画面で「表示」を選ぶと出てきます。
【受講者B】あ、これでできるんですね。そもそもこれ自体はどこから出したんですか？
【福田】Excel の中に入っている Power Query です。Excel 経由で開ける「Power Query エディター」という機能ですね。
【受講者B】なるほど、わかりました。

【後半：例からの列機能・AIによるクレンジングと注意点】

【福田】皆さん、とりあえず条件列の追加でのクレンジングはお試しいただけましたでしょうか？完全には分からなくても「なんとなく分かった」となっていただければと思います。



もう一つ、列の追加からのクレンジングで「例からの列（Column from Examples）」というのをやってみたいと思います。表記が揺れているところを対象にやってみたいと思います。
「例からの列」を押すと、チェックが入っている列を対象にして新しい列を作っていきます。
例えば「モバイルバッテリーを買っておきたい」を「モバイルバッテリー」と打ちます。すると下の方に、水と簡易トイレの備蓄等がある中で、一つ例を入力するとそれを推測して他の行のことも勝手に書き換えてくれるんですね。

や...	今後、ご家庭で新しく始めたいと思われ...	区切り記号の前のテキスト
	モバイルバッテリーを買っておきたい	モバイルバッテリー
	特になし。家具転倒防止対策	特になし。家具転倒防止対策
	水と簡易トイレの備蓄を見直す	水と簡易トイレの備蓄
	今のところ特になし	今のところ特になし
	ローリングストックを実践したいです	ローリングストック
	避難ルートを歩いて確認したい	避難ルート
	非常用持ち出し袋の中身を点検します	非常用持ち出し袋の中身
	ペット用の備蓄をはじめる	ペット用の備蓄
	ハザードマップを家族で確認する	ハザードマップ
	特になし（家具の固定は考えたい）	特になし（家具の固定は考えたい）
	水とトイレの備蓄の見直し	水とトイレの備蓄の見直し
	今のところは特にありません。	今のところは特にありません。
	非常食（保存食）を増やす	非常食（保存食）

これでOKとすると、136個あった選択肢が88個に減りました。
一回作っちゃうと細かいチューニングはできないんですけど、さっきみたいにポチポチやっていくよりはかなり時間を削減できる可能性もあります。皆さんもお手元でやってみてください。
バージョンによってたまに全部 null になったりするので、パワークエリさんと相談してやってみてください。なんとなくいけてるという人は手を挙げていただけますか？あ、いけてますね。ありがとうございます。
ちゃんとやらなくて大丈夫です。「なんとなくこういう作業が必要になるんだな」と認識いただければ全然問題ないです。



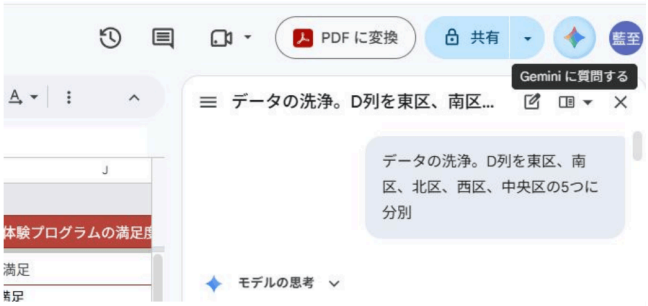
で、身も蓋もないんですけど、手法3として「AIにやってもらいましょう」。皆さんやりたくないですよね、さっきみたいな作業を延々とやりたい人っていないと思うんですよね。細かい作業って本当つまらないので。
「最初からそれ言ってくれればいいじゃん」ということなんですけど、皆さんにデータ洗浄の作業やご苦労を分かっていただいた上で、「やっぱりデータクレンジングって大事なんだな」「楽をするためにはどうすればいいんだろう」という考え方をしてもらいたいということなので、ご理解いただければと思います。

データクレンジングの実践

手法3：AIにやってもらう

GoogleスプレッドシートでGeminiによる処理

自然言語による指示（ノーコード）



AIにやってもらうには、普通に自然言語で指示をすればいいんですよね。例えばさっき条件列を作ってやりましたけど、チャット欄に「データの正常化で5つに分類して」とやると、適当にやってくれます。「中央」とか「南」とか指定しないで単に「5つに分けて」と言ったんですけど、AIが考えて「中央・南・北・西・東」で分けようと分かったみたいですね。綺麗に変わりましたね。

データクレンジングの実践

手法3：AIにやってもらう

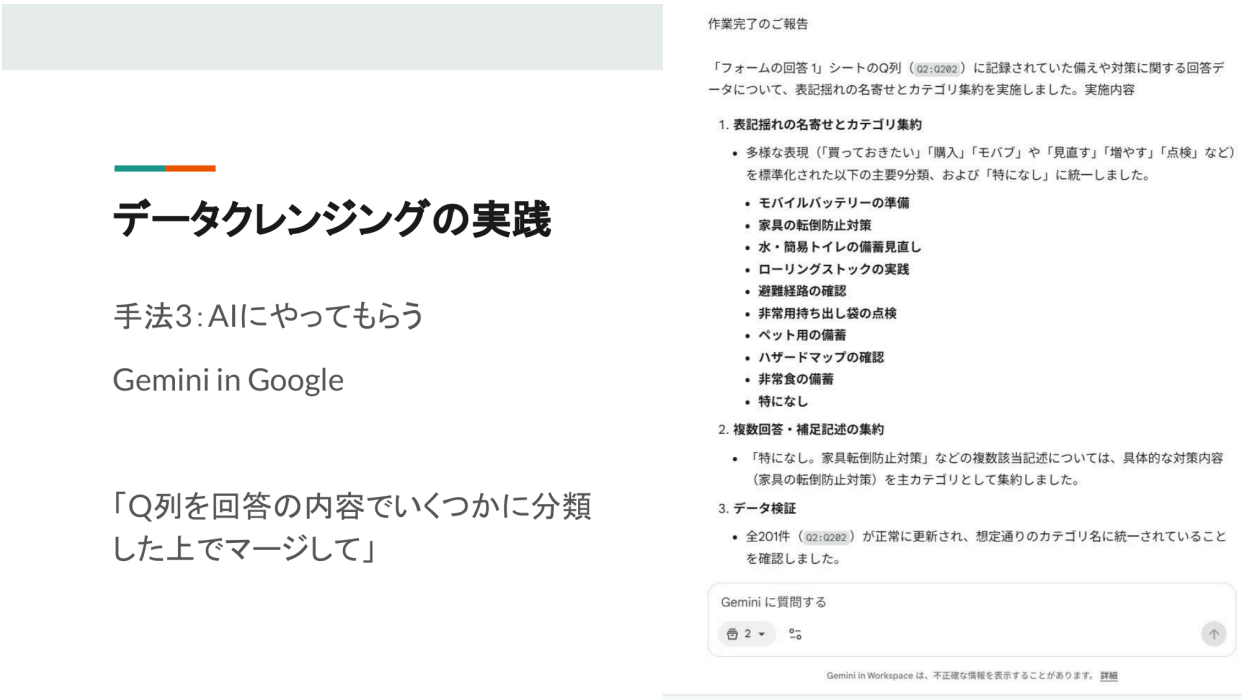
Gemini in Google

「H～N列をとても満足、満足、普通、不満、参加してないの5つに分別。nullはすべて「普通」に置換。複数回答の場合は評価が最も高いものに置換。」



あとプログラムの満足度、混ざったり null が入っていたりするのを分類して良い方に合わせて、null は全部「普通」として置き換えてという指示ですね。そうすると同じような感じで AI が綺麗に並べてくれました。





自由記述の方も、ちょっと乱暴な指示を出しました。「回答の内容をいくつか分類した上でマージして」と。「マージ」という言葉の意味を AI が考えて、色々方法を出してくれました。「表記揺れを直したい」という意味でマージという言葉を使ったんですけど、「振り分けて複数回答は単一回答に集約して

ください」とすると、ある程度カテゴリーをまとめて答えを整えてくれました。

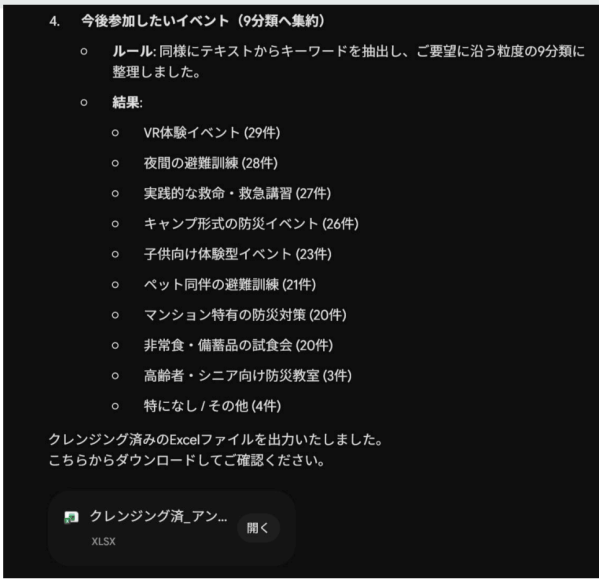


別の方法として、Excelのファイル自体を Gemini 等の AI に突っ込んでみます。添付した上で「中身を見てみて」と指示を出すと、何も言わなくても問題点を全部挙げてくれます。
「このシートダメだよね」「表記揺れがあるよね」というのを何も言わなくても見つけてくれました。このデータ自体は AI に作ってもらったダミーデータなんですけど、タイムスタンプまで見て「これ不自然だよね、何分おきかに投稿されてるし」と診断してくれました。
その上で「どうやって直せばいいの？」と聞くと、直してほしいところを適当に伝えるだけで、最後にはファイルまで出してくれました。見てみると綺麗になってますよね。

データクレンジングの実践

手法3：AIにやってもらう

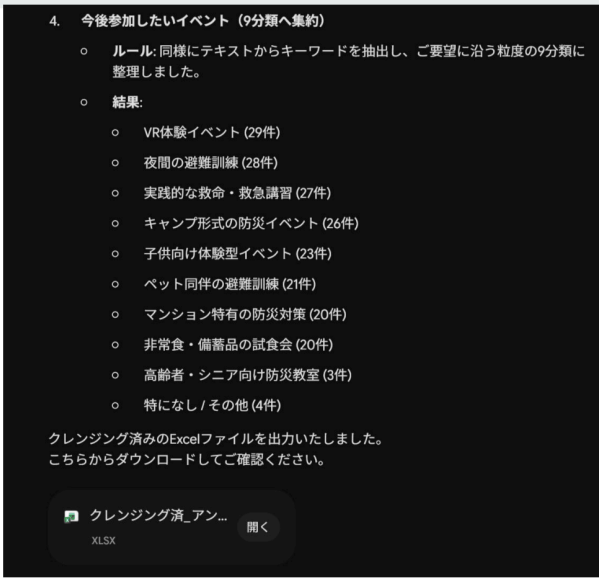
「お住まいの地区」は5分類、プログラムの満足度は5分類、複数回答は評価の良いものに置換、nullは「普通」に置換。今後始めたい備え / 今後参加したいイベントは妥当な6〜9分類にまとめて



データクレンジングの実践

手法3：AIにやってもらう

「お住まいの地区」は5分類、プログラムの満足度は5分類、複数回答は評価の良いものに置換、nullは「普通」に置換。今後始めたい備え / 今後参加したいイベントは妥当な6〜9分類にまとめて



ただ、前回馬場さんの方からもお話ししてもらったんですけど、AIを活用するにあたって注意しないといけないことがあります。

データクレンジング｜ AI活用の注意点

データの秘匿性とセキュリティ：個人/機密情報の入力による情報漏洩や規約違反

→ 事前スクリーニング

ハルシネーション：
存在しないデータへの書き換え、残すべきデータを消去・改ざん

データの保持と品質の劣化：データが壊れるリスク、整合性・正確性の棄損

→ 人によるチェック、バックアップ

収集したデータがどういう属性なのか、十分に注意してください。本来外に出してはいけないデータ、個人情報が入っていたり機密情報が入っていたものを AI に丸投げするのはよろしくないですよね。突っ込むのであれば、事前に個人名や電話番号など直接処理に必要のないところは全部手作業で消去して、スクリーニングしてから突っ込む作業が必要になります。
あとはハルシネーション（幻覚）ですね。データクレンジングにおけるハルシネーションは、データが勝手に消されたり改ざんされてしまう危険があります。存在しないデータへの書き換え、データの品質が低下するかもしれないので、人の手によるチェックや、データが毀損されることに備えて元データのバックアップを取っておくことですね。
Power Query を使うと元データを非破壊処理できるので、バックアップを兼ねた処理としていい方法だと思います。

ヒューマン・イン・ザ・ループ（HITL）

- AI処理プロセスの中に人間が介在する
- 「評価」「修正」「最終承認」は人の役目

AIを使うにあたってすごい重要な概念で「Human-in-the-loop（ヒューマン・イン・ザ・ループ）」という言葉があります。必ず人がチェックする。AIに仕事を任せるんだけど、それを評価したり直したり最終的にOKを出すのは必ず人間がやらないといけない。これは絶対忘れないでください、めちゃめちゃ大事です。

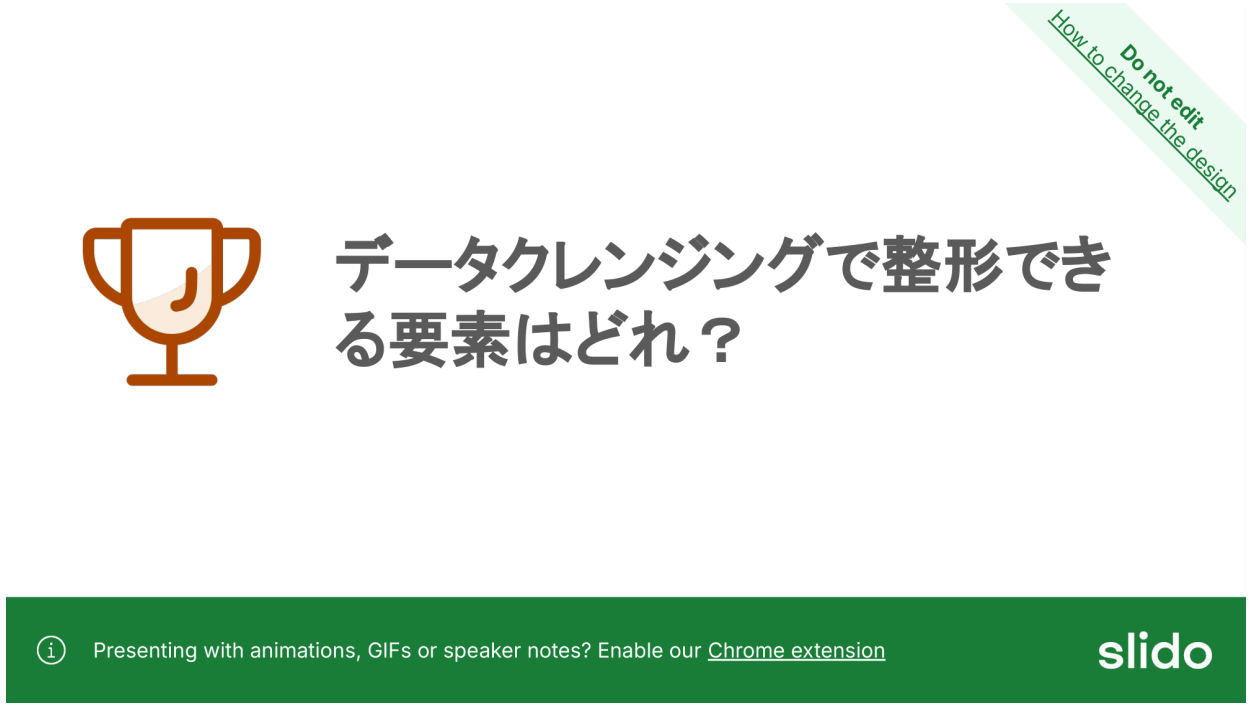
【データクレンジングの限界とクイズ】

データ品質｜6つの基本

① 正確性	データが事実に則しているか	❌ 顧客の引越し前の古い住所がそのまま残っている	⭕ 実際の住所とシステム上の住所が完全に一致している
② 完全性	必要なデータが欠けずに揃っているか	❌ 必須入力であるはずの「メールアドレス」が空欄（NULL）になっている	⭕ 分析に必要な項目がすべて埋まっている
③ 一貫性（整合性）	複数のシステム間でも矛盾がないか	❌ 社内システムAとシステムBで、同じ顧客の電話番号が違っている	⭕ どのシステムを見ても同じデータが同期されている
④ 適時性（最新性）	必要な最新データが利用できるか	❌ 昨日の売上データを見たいのに、反映されるのが3日後になる	⭕ 意思決定をしたいタイミングで、常に最新の鮮度のデータがある
⑤ 有効性（妥当性）	決めたルールや形式に合格しているか	❌ 年齢の欄に「-5歳」や「1000歳」といった値が入っている	⭕ 定められた桁数、日付フォーマット、選択肢（マスタ）の範囲内である
⑥ 一意性	データに重複がないか	❌ 同じ顧客が「会員ID：001」と「002」で二重登録されている	⭕ 1つの実体（人やモノ）に対して、データも1件だけ存在する

データクレンジングでも限界がある

【福田】データクレンジングの実践を通してダメ版のアンケートを綺麗にしたわけですが、限界もあります。
データの品質の6つの基本要素の中で、クレンジングによってある程度整ったけれど洗浄しきれない部分はあるのかなということで、クイズをやります。シートのリンクからクイズスライドを開いてみてください。ニックネーム（偽名で大丈夫です）を入れて参加してください。



（受講者参加確認後、クイズ開始）
【質問1】「データクレンジングで整形できる要素はどれでしょうか？」

- 正解：正確性、一貫性（整合性）、一意性（唯一性）
- 解説：ある程度情報が歪んでいても補正でき、規則に合っていないものも揃えられる。現象として分かれていたものをまとめたりできる。

【質問2】「データクレンジングで補えない（整形できない）要素はどれでしょうか？」

- 正解：完全性、適時性、有効性（※条件付き）
- 解説：
  - 完全性：さっきのケースで言うと未回答のところを「普通」として流しましたよね。本来はそうじゃない可能性もあります。その可能性を全部無視して「普通」に揃えるということは、完全性に欠陥があるデータということです。収集の段階でデータが取得できないのであれば、どんだけ整形しても埋めることができない要素です。
  - 適時性（最新性）：フォームで集めるとタイムスタンプが押されるので問題ないですが、タイムスタンプがないとしたら「いつ取得したものか分からない」場合は埋めることができません。
  - 有効性：例えば年齢でマイナス何歳や千歳は存在しないので「10代～80代以上に収まるように」と人間が定義付けてあげるとある程度担保できますが、機械的にはできません。

完全性は絶対埋めることができないというのを理解していただければ十分でございます。

【構造化データとフォーム設計の重要性】

構造化データ

6つの品質指標（正確性・完全性・一貫性・適時性・有効性・一意性）を満たし、
列と行で整理され、BIツールやAIでそのまま分析できるデータ

↔ 非構造化データ
例）メールやWord・PDFなどの文書、写真、画像、動画、音声ファイルなど

世の中に存在するデータのうち、
構造化されているデータはわずか10～20%

【福田】品質の指標を満たして別行で整理する（スプレッドシートのような形）、BIツールや AI でそのまま分析できるデータを「構造化データ」と言います。クレンジングされたデータ＝構造化データ（整ったデータ）という理解で問題ないです。
逆に構造化されていないデータは画像、動画、音声ファイルなどが含まれます（非構造化データ）。
世の中に存在するデータのうち構造化されているデータは1割～2割程度と言われています。

「あとから構造化」は地獄

- **膨大な手間・時間・人的コスト**
- **完璧な構造化には限界がある**

今回身をもって分かったと思いますが、「後からクレンジングして構造化すればいいじゃん」というのは地獄の作業です、やめましょう。

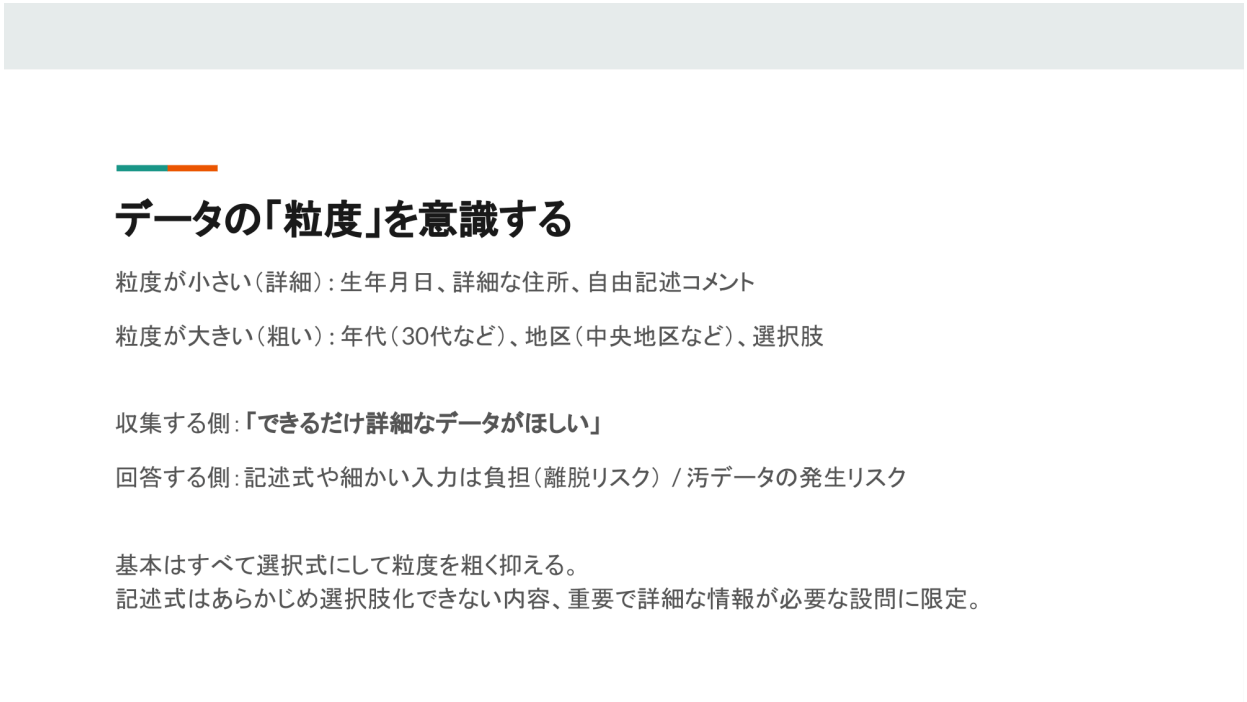
構造化は「収集」の段階で

最良のデータクレンジングは、
最初からクレンジング不要な形（構造化済み）で集めること

→ **フォームの設計が重要**
（ラジオボタン化・回答必須設定など。記述形式は極力除外）

どうするか？「集める段階で綺麗に構造化されたデータになるようにする」、これが一番です。一番いいのはクレンジングしなくていい形で集めること。そのためにはフォームを作って収集する設計が何

より大事です。



一歩踏み込んで説明すると「データの粒度」を意識してください。データの細かさですね。粒度が大きいと大まか（例：20代、30代）、粒度が細かいと詳細（例：28歳、31歳）。
集める側は細かいデータが欲しいんですが、答える側にとっては粒度が荒い投げかけの方が回答しやすいんですよね。より細かいデータを集めようとして記述式や複数選択にすると、答えるのが面倒になって回答してくれないリスクもありますし、記述式に潜むリスクは「データが汚れて整理が難しい」ことです。
なのでフォームを作って情報を集めるのであれば、基本的には「すべて単一回答の選択式」にする。できるだけ粒度を荒く抑える、これが大事です。
どうしても必要な情報、あらかじめ選択肢が作れない本当に重要な情報のみ「記述式」をあてる。基本は使わないけれど、あえて必要なところで使ってください。このあたりを抑えると良いフォームが作れます。

皆様が作っていただいたフォームの例を見ると：

皆様の作例から



- 良い例（PC申請フォーム）：全部回答必須になっていて空白が発生しない。カレンダーや選択肢を使って指定している。時刻のフォーマットを設定して半角全角の表記揺れを防いでいる。

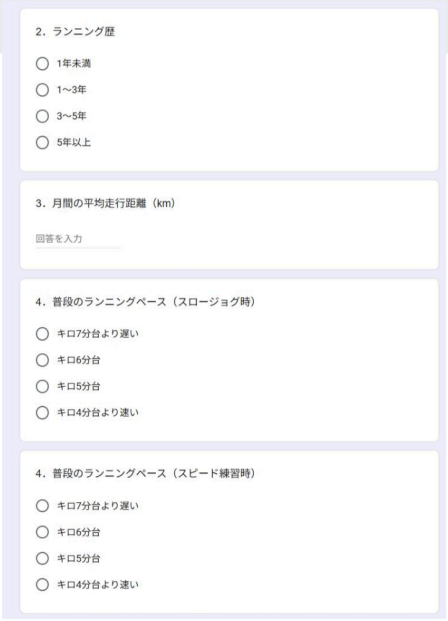
皆様の作例から



- 改善の余地がある例（金融機関名の入力）：銀行名を記述式にすると「青森みちのく銀行」の場合だけでも「青森みちのく」「青森銀行」「青みち」「みちのく銀行」「青銀」「みち銀」等で表記揺れが想定される。選択肢（青森みちのく銀行、青い森信用金庫、東奥信用金庫、その他等）

にすると綺麗に集められる。

皆様の作例から





- センスが良い例（ランニング同好会アンケート）：基本は選択式にしておいて、一番知りたい「走行距離」だけをあえて自由記述の実数で答えさせている。

本日のゴール｜
データ整形の重要度を正しく認識する

長々とお時間かけてご説明しましたが、データクレンジングの重要性についてお分かりいただけたで

しょうか。

【次回（第3回・第4回）のアナウンス】

【福田】第3回は、皆さんがフォームで作って集めたデータをデータポータルで可視化する作業をやっていただこうと思っています。10月中にフォームでデータ収集をしておいてください。
第4回（最終回）はミニプレゼンをやっていただきます。第3回で作ったデータポータルの画面を使って、集めた情報を可視化した上でサマリーを説明していただきます。庁内向けなら「このフォームでこういうデータが集まって業務効率化できる」、庁外向けなら「こういう声が集まったのでこういう施策を展開する」といった意思決定を促すようなプレゼンです。
宿題として時間外でやってもらうのは極力避けたいため、第3回の中でデータポータルの仕込みとプレゼン準備を一緒に進められればと思っています。
第1回に参加できなかった方やフォームをまだ作っていない方は、10月中に無理なくデータ収集できるフォームを作って第3回に備えてください。フォローが必要な方は私（福田）までメールでお知らせください。
fukuda@consis.jp

お時間になりましたので、第2回分科会を終了とさせていただきます。長時間お疲れ様でした。末尾のアンケートにご協力をお願いいたします。