教職課程センター紀要　第８号　47-51 ページ、2023 年 12 月
Jour. Center Teacher Develop. Edu. Res., Daito Bunka Univ., No.8 47-51, 2023
論文

データサーエンスとミュージアム

Data Science and Museum

和田　浩
Hiroshi WADA

Key words: Museum, Data Science, Statistics Analysis

Museums collect a variety of data in the course of their daily operations. For example, data such as temperature and humidity of exhibition rooms and the number of visitors are examples of data collected in many museums. And many formats are used to store the collected data, with the main formats being printed on paper, Excel data format, pdf format, text format, etc. These data cannot create much value simply by being stored, and it is only when the data is utilized that it becomes "information" with value as an asset. It is no exaggeration to say that there is unlimited potential for future utilization of the data accumulated by museums, many of which are in a dormant state or have not been fully utilized. In order to maximize the potential of the data, it is of course a prerequisite that the data be ready for utilization. Ideally, data should be put into a form that is easy to utilize at the time of collection, but the reality is that means must also be devised to utilize what has already been accumulated in the field. This paper takes the latter viewpoint and discusses the processing of data on the number of people staying in an exhibition room as an example.

１．はじめに

ミュージアム（博物館や美術館等の施設）では日常業務の中で様々なデータが収集されている。例えば、展示室の温湿度、入場者数といったデータは多くのミュージアムで収集しているデータの一例として挙げられる。そして、収集したデータの保管には多くの形式が採用されており、紙にプリントした形式、エクセルデータ形式、pdf 形式、テキスト形式等が主な形式となる。これらのデータ群は保管されているだけでは多くの価値を生み出すものとはなりえず、データを活用することで初めて資産としての価値を持つ「情報」となる。ミュージアムが蓄積しているデータには、眠っている状態のものや、十分に利活用されていないものも多く、今後の利活用には無限の可能性が存在すると言っても過言ではない。その可能性を最大限に発揮するためには、もちろんデータが利活用できる状態に整っていることが前提となる。理想的にはデータ収集時点で、利活用しやすい形式に整えられるべきであるが、現場で既に蓄積されたものを活用する手段も一方では考案せねばならないのが実状である。本稿は後者の視点に立ち、展示室における滞留者数のカウントデータの処理を一例として論じるものである。

２．データ収集時における注意事項とは

データに対して統計的な処理を施す場合、処理対象となるデータがどのようなフォーマットで保存されているのかは大変大きな意味を持つ。適切なフォーマットであればいかなる統計処理であっても直接適用できるため、処理が迅速に進む。一方、不適切なフォーマットであれば、統計的処理を施す前に、フォーマットを整える作業が生じてしまう。こうした作業は「前処理」や「データクレンジング」と呼ばれ、実はデータサイエンスの工程で多くの労力が費やされる工程でもある。もしも、現在収集しているデータのフォーマットに何らかの不備が存在するのであれば、然るべきタイミングでフォーマットを改善しておくと、将来的に非常に有益な効果をもたらすことは間違いない。

このような、データ収集時における基本的注意事項として、大変参考になるのが、令和 2 年 12 月 18 日付の総務省報道資料「統計表における機械判読可能なデータの表記方法の統一ルールの策定」で提示された資料である

[1)]。重要と考えられる部分を同資料別紙より抜粋すると、「１セル１データとなっているか」「数値データは数値属性とし、文字列を含まないこと」「セルの結合をしていないか」「スペースや改行等で体裁を整えていないか」「数式を使用している場合は、数値データに修正しているか」「１シートに複数の表が掲載されていないか」といったチェック項目が実例とともに列挙されている。

では、一体なぜこうした注意事項が挙げられたのかを考えると、おそらくエクセルに代表される表計算ソフトを、統計処理よりも文章編集を目的として使用する局面が多い実状がまず原因していると推測される。当初はデータ入力フォーマットとして作成したものが、そのままデータ処理結果のプレゼン資料や配布資料として使用されることは多くのミュージアム職員が経験していることであろう。あるいは、レイアウト調整しやすいエクセルのシートがワードに代替するものとして用いられる事例も多々見受けられる。使用法は各自の自由であるし、１つのファイルで複数の用務をこなすことは効率的ではある。しかしながら、特定資料の１回の用途だけに使用され、以降のデータの利活用は考慮されないのであれば、再度同じデータを異なるフォーマットで収集するか、加工するといった手間が必ず生じてしまう。統計処理をするために、同じ数値をコピー・ペーストして別のエクセルファイルを作成するといった作業は、結果的にはトータルで考えると多くの時間と労力を消費してしまうことになり、実は非効率な結果を生み出すことに繋がってしまう。また、元データに誤りが存在していた際、元データが更新されたタイミングで、データを引用した各種ファイルが自動的に更新されず、全て手作業に依存する事態に陥ってしまう。データ量が少なければ何ら問題は生じないが、膨大なデータである場合、あるいは経営に対して非常に深刻な影響を及ぼすデータ、または、公開に際して慎重を要する類のデータである場合には取り返しのつかない混乱と問題が生じる可能性がある。つまり、データを利活用するフレームワークの設計が極めて重要なのである。

３．ミュージアムが混雑する時間帯は？

2022 年 5 月 3 日から 6 月 26 日の間、東京国立博物館は『沖縄復帰 50 年記念　特別展「琉球」』と題した特別展を開催した（以降、琉球展と表記する）。琉球展はコロナ禍が完全に収束する前の時期であり、かつ多数の来場者が見込まれたが、時間帯指定の予約制は採用せず、チケットが販売された。琉球展では、30 分毎に特別展会場に滞留している来場者数（滞留者数）を手作業でカウントしており、そのデータが東京国立博物館に保存されている。通常の特別展では会場入口のもぎりで、30 分毎の入場者数をカウントしており、琉球展ではそのデータも保存されている。周知のとおり、コロナ対策において、いわゆる 3 密の回避が多くの施設に求められ、もちろんミュージアムもその例外ではない。良好な換気環境を維持しているかを判断する尺度として、空気中の二酸化炭素濃度が用いられる。二酸化炭素濃度は、空間内に滞留する人間の数と相関がある。滞留者数が多いと、空調機器を用いて積極的に換気、すなわち新鮮な空気である外気を取り入れる必要が生じる。しかし、空調機器の能力以上の換気はできないため、許容可能な滞留者数は施設のスペックによって概ね定まる。つまり、ミュージアム施設がどの程度の滞留者数を許容できるのかを知るためにも、現実の滞留者数は必要なデータであり、上記のようなコストを費やしてカウントをしたものである。本稿における狙いとしては、コストを費やして獲得した滞留者データを、単に蓄積するだけではなく、そこから新たな価値を創造できるような情報を抽出する点にある。これらのデータを統計処理し、琉球展の会場の混雑度合いを可視化することとした。

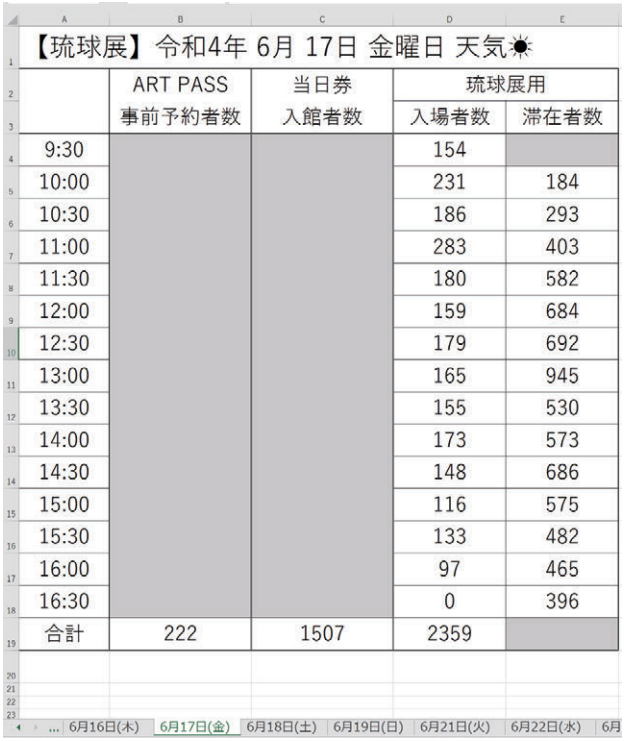
４．データの前処理

統計処理を効率的に実施するためには、データの前処理を要する。まずは、エクセルで作成された琉球展の滞留者数データの形式（図 1）を確認すると、

- 1 つのブックに複数のシートが存在する。
- 各シートがある日のデータとなっている。
- 各シート名は日付曜日（例「6 月 17 日（金）」）となっている。
- 1 行目から 3 行目には、列の名称や天気が記載され、統計処理には不要。
- 最終行には B 列、C 列、D 列、E 列の合計値が記載されている。
- A 列から E 列の 5 つの列が使用されている。
- A 列は時刻が文字列で記載されている。
- A 列は「9:30」から 30 分間隔で「16:30」まで記載されている。
- B 列は事前予約者数、C 列は当日券入館者数であり、ともに最終行にのみ合計値が記

載されているのみ。
・D 列は入場者数、E 列には滞留者数が記載されている。

というものである。



【琉球展】令和4年 6月 17日 金曜日 天気☀

	ART PASS 事前予約者数	当日券 入館者数	琉球展用	
			入場者数	滞在者数
9:30			154	
10:00			231	184
10:30			186	293
11:00			283	403
11:30			180	582
12:00			159	684
12:30			179	692
13:00			165	945
13:30			155	530
14:00			173	573
14:30			148	686
15:00			116	575
15:30			133	482
16:00			97	465
16:30			0	396
合計	222	1507	2359	

6月16日(木) 6月17日(金) 6月18日(土) 6月19日(日) 6月21日(火) 6月22日(水) 6月

図 1 日常的な記録に用いられていたエクセルフォーマット

そこで、統計処理を効率的に行える形式に変換するためには、以下のような手順による前処理が必要であると考えた。

・B 列と C 列は不要のため削除する。
・A 列を年月日時刻「yyyy/mm/dd hh:mm」の形式にする。
・ブック内の各シートを縦連結する。
・連結後のものを csv ファイルで保存する。

細かな部分を追記すると、

・シートの名称から月日情報を抜き出し、それを月/日形式に変換する（例：「6 月 17 日（金）」を「6/17」に変換）。
・シートには年の情報が存在しないので、縦連結する前に、年の情報をユーザーが入力する。
・ユーザーが入力した「年」、シート名称から作った「月/日」、A 列の「各時刻」を合成して年月日時刻「yyyy/mm/dd hh:mm」の形式を作る（例：「2022」と「6/17」と「9 時 30 分」から、「2022/6/17 09:30:00」を作る）。
・A 列は展覧会初日の午前 0 時 0 分から始まり、最終日の午後 23 時 30 分までの 30 分間隔の連続データとする。
・元シートの A 列に無い年月日時刻データは自動的に作成して埋める。
・D 列および E 列のデータが存在しない箇所は 0 とする。

といった処理が含まれる。これらの処理は手作業ではなく Python によるコードを作成し、対象のエクセルファイルを指定すれば一瞬で連結後の csv ファイルが生成する仕組みを構築した。Google のアカウントを持っていれば無料で使用可能な Google Colaboratory 上で動くように作成した[2)]。以上の前処理を施した結果、非常にシンプルな csv ファイルが生成された（図 2）。

	A	B	C
1		entry	stay
2	5/3/2022 0:00	0	0
3	5/3/2022 0:30	0	0
4	5/3/2022 1:00	0	0
5	5/3/2022 1:30	0	0
6	5/3/2022 2:00	0	0
7	5/3/2022 2:30	0	0
8	5/3/2022 3:00	0	0
9	5/3/2022 3:30	0	0
10	5/3/2022 4:00	0	0
11	5/3/2022 4:30	0	0
12	5/3/2022 5:00	0	0
13	5/3/2022 5:30	0	0
14	5/3/2022 6:00	0	0
15	5/3/2022 6:30	0	0
16	5/3/2022 7:00	0	0
17	5/3/2022 7:30	0	0
18	5/3/2022 8:00	0	0
19	5/3/2022 8:30	0	0
20	5/3/2022 9:00	0	0
21	5/3/2022 9:30	299	0
22	5/3/2022 10:00	121	341
23	5/3/2022 10:30	136	396
24	5/3/2022 11:00	180	431
25	5/3/2022 11:30	114	543
26	5/3/2022 12:00	111	639
27	5/3/2022 12:30	153	433
28	5/3/2022 13:00	133	402
29	5/3/2022 13:30	157	435
30	5/3/2022 14:00	139	477
31	5/3/2022 14:30	177	481
32	5/3/2022 15:00	148	483
33	5/3/2022 15:30	119	488
34	5/3/2022 16:00	64	438
35	5/3/2022 16:30	6	367
36	5/3/2022 17:00	0	0
37	5/3/2022 17:30	0	0
38	5/3/2022 18:00	0	0
39	5/3/2022 18:30	0	0

図 2 前処理後のデータの一部

５．統計処理

おそらく、一般的には各日に計測した滞留者数を折れ線グラフで描画する可視化は実行されていると想像する。しかし、現実的には多数の折れ線グラフを用いて、データ群から特徴を抽出するのは非常に難しい（図3）。

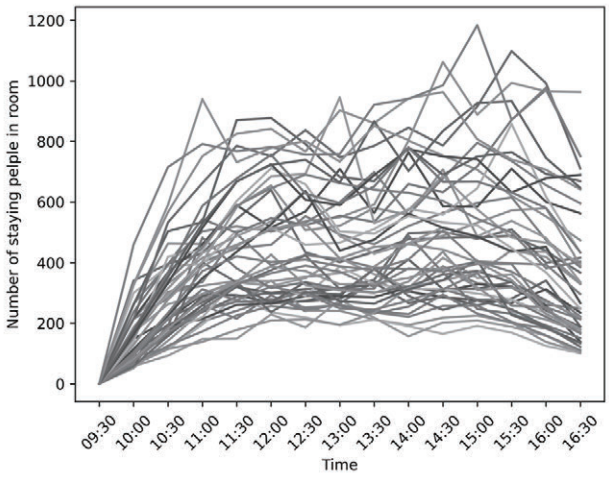


図 3 滞留者数の折れ線グラフ（X 軸：時刻、Y 軸：滞留者数）

そこで、ひと手間加えた解析を試みた。具体的には、日付に関係なく各時刻における滞留者数のばらつきを見るというものである。例えば、11 時における滞留者数としては何人くらいが最も高頻度で計測されたのか、といった視点で解析を行なうと、日によってばらつきはあるものの、1 日の内で最も滞留者数が多い時間帯や、少ない時間帯を導き出せる可能性がある。つまり、時間経過に伴う滞留者数の増減に一定の傾向が存在するのかどうかを判別できることになる。こうした傾向が分かれば、比較的空いた時間帯に展覧会を鑑賞したい場合、何時頃にミュージアムへ行くのが希望通りの環境に遭遇する確率が高いのかを予測できる。施設の経営者側にとっては、さらに集客が見込める時間帯や、お出かけを予定しているお客様への混雑予想の提供といったカスタマー向けサービスの展開といった部分に有益な情報をもたらすのではないかと考えた。

まず、同じデータを用いて、時間帯毎の滞留者数についてヒストグラムを作成した。ヒストグラムを 3D で描画すると、棒が高い部分が最も頻度が高いことを意味する（図 4）。つまり、琉球展では、棒の高い部分の滞留者数が、その時刻における滞留者数として多く観測された値ということになる。この 3D ヒストグラムは棒の高さに応じて色が変化するフォーマットに仕立てている。

したがって、上方から垂直に見下ろす視点で 2D のヒストグラムを描画すると（図 5)、棒の色のコントラストがより識別しやすくなる。これらの解析から、琉球展においては展示室内の滞留者数は「M」型の変化傾向を持っていることが分かった。具体的には、開館直後から滞留者数は上昇し、12 時頃にピークに到達する。以降は減少し、13 時頃から再び上昇する。15 時頃に 2 度目のピークに到達し、閉館時刻まで減少する、といった傾向である。開館直後を除外すると、混雑度合いが低いのは 13 時頃と 16 時 30 分頃という結果となった。

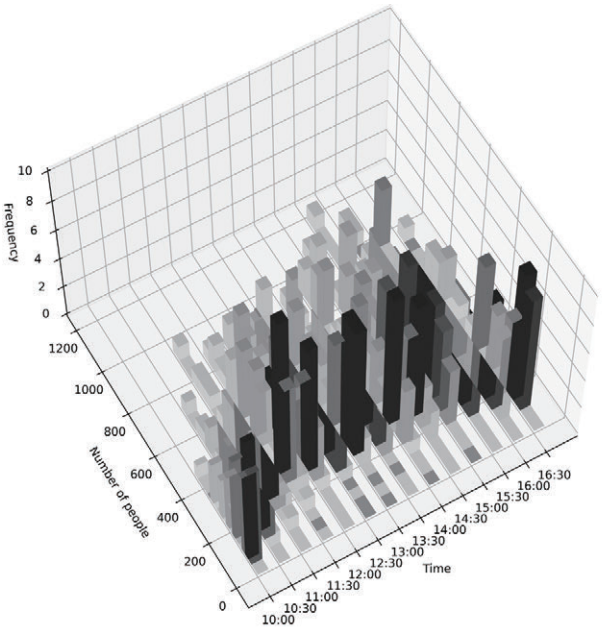


図 4 時間帯毎の滞留者数の 3D ヒストグラム（高さ方向が頻度の軸であり、棒の色の濃さが対応する）

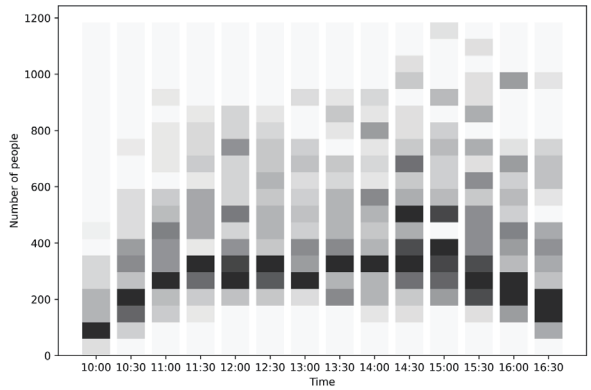


図 5 時間帯毎の滞留者数ヒストグラムを平面化したもの（X 軸：時刻、Y 軸：滞留者数、図 4 の棒の色濃度を反映）

６．ミュージアムにこそデータサイエンスを

本稿で提示したのは、簡単な統計処理の一例であるが、同じデータであっても、加工方法によって表現や得られる解析結果の深みに違いが出ることは明らかである。ミュージアムには日々の活動の結果蓄積された様々なデータが存在する。それらを眠らせておくだけでは非常に勿体ない。データを深堀りすることで、予想外の成果が得られることもある。その成果はデータのもつ新たな価値であると言い換えることができる。つまり、ミュージアムには価値のあるデータが既に多数存在しており、データサイエンス技術によってその価値を見出すことが結果的にミュージアムにとっての利益として還元されると考えられる。文系施設としての色が強いミュージアム業界ではあるが、今こそデータサイエンス技術によってミュージアム自体が新たなステージに進むことのできるタイミングではないだろうか。

引用文献

１）https://www.soumu.go.jp/menu_news/s-news/01toukatsu01_02000186.html（令和 2 年 12 月 18 日付 総務省報道資料「統計表における機械判読可能なデータの表記方法の統一ルールの策定」

２）https://github.com/hiroshiwada2020/2023daito.git

謝辞

本研究は JSPS 科研費 23H00024「文化遺産アセットの効率的利活用を目指したミュージアム DX 技術の開発（研究代表者：和田浩)」の助成を受けたものである。