データ分析における問題定義や目的設定の有無によって生じる分析過程の違い

Differences in analytical process caused by problem definition and purpose setting in data analysis

辻 泰輝[*1], 山崎 治[*1]
Taiki TSUJI[*1], Osamu YAMAZAKI[*1]
[*1] 千葉工業大学情報科学部
[*1]Faculty of Information and Computer Science, Chiba Institute of Technology
Email: used.mmm.10@gmail.com

あらまし: 本研究の目的は，事前の問題定義や目的設定の有無がデータ分析の過程や結果の洞察に及ぼす影響を調査することである．実験では「目的あり群／なし群」に分けた参加者を 2 人ずつの組にし，仮想的な店舗の売り上げデータに基づいた複数のグラフを提示した上でデータ分析を行わせた. 分析過程について参照されるグラフの推移を整理した結果，両群の間に目的の有無が及ぼした差として，データを示すグラフの閲覧・利用の仕方に違いが現れることがわかった.
キーワード: データ分析，問題定義，目的設定，データアプローチ，課題アプローチ

1. はじめに

ビッグデータを意思決定の材料として活用するために，データ分析を活用したいと考えている企業は多く存在し，ビジネス現場での利用価値への期待は高まっている．しかし，目的通りに活用出来ていない企業が散見される．

データ分析を行うことで何かしらの結果は得られたものの, その結果が新しい発見に繋がるわけでも，ビジネスチャンスへ繋がるわけでもない．このような状況に陥ってしまう原因の 1 つとして考えられるのが，「データアプローチ」という考え方である(1). 「データアプローチ」とはデータありきで，その活用方法を導き出そうとする考え方のことである．本来データ分析というのは，現状抱えている課題の把握や，その課題を解決するために必要なデータや，分析の目的をまず明確に定める必要がある．その上で，データを収集し，分析を行うという流れが正しいデータ分析の考え方である. このような考え方を，河村ら(1)は「課題アプローチ」と呼んでいる．

2. 目的

本研究では，「データ分析の目的」に焦点を当て，目的を持ってデータ分析を行った場合と，そうでない場合での分析結果に差が現れるのかを調査し，それらの重要性を明らかにすることを目的とする．そこで実際のデータ分析を通じて考察を行う課題を利用した実験を行う．特に，目的設定に関して，実験者側から介入を行う参加者グループと，介入を行わないグループの違いに注目し，分析の過程に違いが表れるかを分析する．

3. 実験　目的の有無がデータ分析に及ぼす影響

実験では、目的のあり／なしが分析過程や結果の洞察に影響を及ぼすのかを調査した．

3.1 方法

実験参加者：情報科学もしくは経営工学を専攻とする大学4年生12名が2人1組ごとに実験に参加した．計 6 組の参加者を，目的なし群（3 組），目的あり群（3 組）に分けた．

実験計画：1 要因 2 水準参加者間計画で行う．独立変数として目的の有無を取り上げ，「目的を与える」／「目的を与えない」の 2 水準を設ける．

材料：Kaggle 社の Web ページに掲載されている「 Store Item Demand Forecasting Challenge（https://www.kaggle.com/c/demand-forecasting-kernels-only/data）」という商品売上のデータを用いた．同データを加工し，仮想的な実験用データとした．実験用データは「月」「日」「曜日」「店舗」「アイテム」「年齢」の 6 項目で構成され, それぞれに対して「売上（個数）」が示される．また資料として， 1 つの項目に対する売上を示すグラフ（単純集計）を 6 個，2 つの項目を掛け合わせたものに対する売上を示すグラフ（クロス集計）を 15 個 Excel 上で作成した．

手続き：実験は「データの観察・話し合い（25 分間）」・「記述（制限時間なし）」の二段階で構成される．目的あり群，目的なし群の共通目的として，実験用データから読み取れる「現状」と，それに対する「改善案」を提出することを求めた．この際，目的あり群のみに「既存商品・既存店舗についての弱みを知りたい」という分析の目的を伝えた．

「データの観察・話し合い」段階では， 25 分間 2 人でデータを観察しながら，気づいた点，考えた内容をメモ用紙として用意した A4 用紙 1 枚に記入させた．また参加者の分析の様子をスクリーンキャプチャソフトで録画した. 25 分後,「記述」段階では，データを閲覧させず，メモ用紙に記した内容をまとめ「現状」と「改善案」の記述を行なわせた．参加

者は，実験者が PowerPoint で作成したフォーマットに「現状」および「改善案」を記入した．

3.2 結果

参加者が閲覧したグラフや，提出された現状と改善案に大きな差は見られなかった．しかし，閲覧したグラフが単純集計であったのか，それともクロス集計であったのかについての割合では，図 1 のような差が見られた．

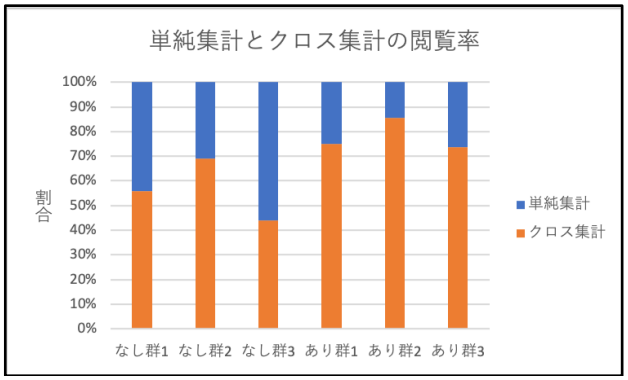

図 1 単純集計とクロス集計の閲覧率

図 1 から目的なし群は 70%以下，目的あり群は 70%以上クロス集計を閲覧した結果となった．本実験では分析資料として単純集計グラフを 6 個，クロス集計を 15 個作成した為，おおよそ 3 : 7 の割合である．この結果から，目的なし群は単純集計を，目的あり群はクロス集計のグラフをより多く閲覧したと読み取ることが出来る．

また，目的なし群，目的あり群別にグラフ 21 個ごとの閲覧数を図 2 に示す．縦軸が閲覧数の平均値，横軸がグラフの番号を示す．グラフ番号の 1〜6 番が単純集計のグラフ，7〜21 番がクロス集計のグラフである．

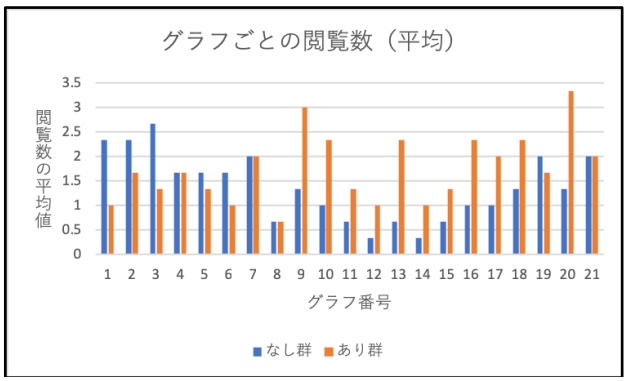

図 2 グラフごとの閲覧数（平均）

目的なし群は単純集計グラフを，目的あり群はクロス集計のグラフを多く閲覧していることが図 2 からも分かる．

目的なし群と目的あり群との間で，閲覧数の差が特に大きいグラフは 9（店舗×月），10（アイテム×月），13（店舗×日），16（店舗×曜日），20 番（店舗×年齢）のグラフである．全てのグラフがクロス集計であり，「アイテム」もしくは「店舗」の項目が含まれるグラフである．

4. 考察

まず，単純集計とクロス集計の閲覧率について，目的あり群は「既存商品・既存店舗についての弱みを知りたい」という分析の目的を与えられていた為に，目的あり群は「アイテム」と「店舗」の項目に着目しやすい．その結果，「アイテム」と「店舗」を他の項目と照らし合わせる為に，クロス集計のグラフを単純集計のグラフより多く閲覧したのではないかと考えられる．

また，参加者がデータ分析にあたり，「分析の軸／観点」を設定しながらグラフを閲覧していたことを確認するため，グラフの閲覧順序について分析した．その結果，目的あり群の方が同じ項目を含むグラフを連続で閲覧する傾向にあった．また，分析過程全体を通じて特定の項目のみを閲覧しているような傾向は読み取れず，目的あり群は単体の項目に対する売上ではなく，データを比較しながら閲覧しようとする姿勢があったのではないかと考えられる．

最後に，参加者が提出した「現状」について，目的なし群と目的あり群との間に差を確認することはできなかった．その原因は，題材として使用した実験用データが簡略過ぎた為であると考えられる．本実験においては「改善案」は分析の対象としなかった．理由としては，どの組も思いついたアイデアを記入していた為に比較が困難であった為である．

本実験では主に参加者が閲覧したグラフに焦点を当て，結果の分析を行った．今後，録画・録音した参加者の協調作業の様子やメモの内容，使用したグラフフィルターの内容等も踏まえた上での分析を行う予定である．特に，録画・録音からは各参加者群の思考過程について，グラフフィルターの使用内容からは各グラフの中で着目していた要素について，より細かく観察することが可能であると考えられる．

5. まとめ

本実験において，現状および改善案の報告では明確な差がみられなかったものの，分析過程においては両群の間に異なる特徴がみられた．

使用したデータや，分析目的の提示方法に課題が残る結果となったが．今後の展開として，「分析で解決したい問題の定義」や，その問題定義に沿った「分析目的」をさらに詳細に設定し，データ分析過程の違いを検討していく必要がある．

参考文献

(1) 河村真一・日置孝一・野寺綾・西腋清行・山本華世(著)・日経情報ストラテジー(編) (2016). 本物のデータ分析力が身に付く本 日経 BP 社出版