技術トレンド

TREND INFORMATION

Pandasを利用したデータの取り扱い

企画連携課 企画連携係

データ活用の重要性については様々な場面で挙げられていますが、いざ取り組むとなると、データの収集・整理・分析と、負担は大きいものです。効率良く作業するためには、データの変形や集約といった柔軟な操作が重要であり、そのための手段の1つとしてpandasの利用が挙げられます。

pandas概要

pandasとは、プログラミング言語Pythonで利用することができる、データ分析を支援するオープンソースライブラリのことです。数多くあるプログラミング言語の中でも、Pythonはコードの記述がシンプルで多様なライブラリが提供されており、機械学習などにも利用されています。

pandasでできることは非常に多岐に渡りますが、大雑把に言うとコマンド操作型のExcelなどをイメージすると分かりやすいかもしれません。マウスを使った直感的な操作ができないため扱いが難しく思えますが、複雑な処理を簡潔に実行できる多様なコマンドが用意されています。大量のデータを扱う場合や繰り返し作業を自動化したい場合など、目的に応じてpandasを利用することで効率的に作業を行うことができます。

時系列データにおける利用例

気象データを例に、pandasのイメージをご紹介します。表1は気象庁で公開されている京都市の過去1年分の気象データを引用したものです。(JupyterLabと呼ばれるPythonの実行環境を利用し、CSV形式のデータを読み込んでいます)

表1 過去の気象データ

	年月日	平均気温(℃)	最高気温(℃)	最低気温(℃)	降水量の合計(mm)	降水量の合計(mm).1
0	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	現象なし情報
2	2022/4/1	8.8	12.6	5.2	0.5	0
3	2022/4/2	8.8	15.8	2.8	0.0	1
4	2022/4/3	10.9	13.6	8.0	0.5	0

例えば、記録保存していたデータがこのような日データで、月毎の集計データが必要となった場面を想定します。

pandasでは時間の要素を持つ時系列データに対して、期間を指定して集計するコマンドが用意されています。注意が必要なのは、時間のデータをコンピュータが時間であると認識している必要があります。例えば表1の場合、現状ではコンピュータには「年月日」のデータが日付ではなく単なる文字の並びとして認識されています。そこで、時系列データとして扱うために「年月日」が日付型であると指定しインデックス(見出し)にします。利用しない行と列を削除することで、表1から表2のようにデータを整形します。(表1の6列目は降水現象の有無を表記)

表2 整形後のデータ

年月日	平均気温(℃)	最高気温(℃)	最低気温(℃)	降水量の合計(mm)
2022-04-01	8.8	12.6	5.2	0.5
2022-04-02	8.8	15.8	2.8	0.0
2022-04-03	10.9	13.6	8.0	0.5

このデータを月毎に集計しようとするのですが、各項目は平均・最高・最低・合計と、集計方法がそれぞれ異なっています。このような場合でもpandasでは列毎に違う統計量を一度に計算することができ、集計期間の指定と組合せて、表3のような集計を1コマンドで実行することができます。

表3 集計後のデータ

年月日	平均気温(℃)	最高気温(℃)	最低気温(℃)	降水量の合計(mm)
2022-04-30	16.530000	28.9	2.8	127.5
2022-05-31	19.716129	33.5	7.7	80.5
2022-06-30	24.430000	37.2	13.9	125.5

全体作業としては、①csvの読込み&書式指定、②不要行の削除、③不要列の削除、④集計の4コマンドで実行することができます。年月日が各月末日となっていますが、表記を月までに変換することも可能です。年度の違うデータや集計期間の変更といった別条件を調べたい場合は、コードの一部を変更することで実行することができます。

活用と気をつけたい点

例は品質の良いデータ例でしたが、実際には重複や欠損などがデータに含まれることはよくある話です。それらをどう扱うかは判断が必要になりますが、pandasでは削除・置換といった操作も容易です。条件別にグループ化して計算したり、共通項目によるデータ結合など、時系列以外にも様々な種類のデータで活用することができます。一方で、コマンド操作でミスに気づかなかったり、操作に慣れるまでの初期学習やPythonの基礎知識などが必要となる面もあります。コスト以上にリターンが得られる可能性がある一方で、状況によっては別の手段を使う方が効率的な場合もあります。目的と状況に応じて、pandasを手段の1つとして検討してみてはいかがでしょうか。

●お問い合わせ先／ 京都府中小企業技術センター 企画連携課 企画連携係　TEL：075-315-8635　E-mail：kikaku@kptc.jp