第3回カンニングペーパー

P2

問7の開設をやっていきます。

問題は、practiceテーブルの各項目について、以下の項目をということで

最大値、最小値、充足率を出力し、不適切な値がないかをチェックしてください。

充足率とは、全体の件数に対する各項目の値が入っている件数割合のことです。

この問題は今後演習課題として出てくるデータのプロファイリングに通じるところがある問題になっているので、ぜひどんな値が出力されるかということに注目して解き進めていってもらえればと思います。

〇最大値

まずは、各項目の最大値を求めます。

最大値を求め際は、集約関数のMAX()関数が使えます。使い方としては

MAX()かっこの中に項目名を入れてやればよいです。

〇最小値

最小値は、先程集約関数としてMAX()関数を使っていたところをMIN（）関数に変えるだけです。最大値と最小値について、項目が文字列の値の場合は辞書順に小さいまたは大きい値が出力されます。

〇充足率

最後に充足率です。充足率は全体の件数に対する各項目の値有りの件数の割合なので、

COUNT()関数を使用し、全体の件数と各項目の件数を計算し割り算してやればよいということになります。

この方法でそのまま実装するとここに書いてある通りで、COUNT()アスタリスクで全体の件数、COUNT項目名で各項目の件数をそれぞれ抽出し、全体の件数で割って算出しています。

ここでCOUNT()も含めた集約関数のポイントとして、COUNT()アスタリスクの場合はNULL値も含めて、COUNT()項目名とした場合は、NULL値は除外して集計されることに注意してください。

一見これで問題ないように思えますが、この方法で出力すると

出力値として、0か１のみしか出力されません。

P.6

この理由には次の2つのことが原因となっています。

1つ目はSQLSrverでは整数同士で割り算をすると、小数点が付くような場合でも切り捨てを行い、整数でしか値を返さない決まりになっているためです。

例？

これは実際に例を見てもらえればわかりやすいかと思いますが

例えば5/2といった計算をSQLserverで実施してみます。

整数のまま記述した場合と、あえて小数点まで記述し、計算すると、このように整数同士の場合は、2、小数点も含めると2.5が返ってきます。

整数同士の場合は暗黙の内にこの小数点以下の値が切り捨てられていることが確かめられるかと思います。

２つ目はとして、今回、件数の取得にCOUNT()関数を使っていますが、COUNT()関数の決まりとして、返ってくる値は必ず整数となっています。

そのため、先程のクエリの場合は1か０かという値しか返ってこないこととなっています。

この対処方法として、ここにある2つの方法が利用できます。

1つ目、整数値に1,0のように小数点を含む数値型の値をかけて強制的に数値型とみなす処理を実施する方法。

２つ目、SQLServerにもともと用意されているデータ型を変換する関数を用いてめいじてきに数値型へ変換するという方法です。

例えばこの一つ目の方法を実装するならばこのようになる。

P.7

ここでは、分子に1.0をかけて数値型に変換したうえで、割り算を実施しています。

P.8

こちらは関数を用いる方法です。

ここでは、CONVERTという関数を用いて、分子をFLOAT型に変換しています。

このようにすることで、1か０かといった返り血でなく具体的に各項目の充足率が何％か求めることができます。

ここまで、最大、最小、充足率を求める方法をご説明しましたが

それぞれの値が求まったら、今回求めた値の中身についてもよく見ていただきたいと思っています。

最大/最小といった極端な値を見たり、充足率といった指標を算出してみることで

データがどういう状態にあるのかを把握できます。

データの定義と見比べながら、何か気づくことがないか、考えてみてください。

解説の続き。

こちら問題文より抜粋した、データの内容説明です。

データは3年生のテスト結果とあります。

しかし今回の調査で、100点満点なのにそれを超える値で登録があったり

他の学年の値で登録されている生徒がいることデータそのものに欠損がることが明らかになります。

データ量が多くなると、すべてのデータの値をみるわけにもいかないので、

こういった調査で、データ品質や状態を確認します。

P.1０

クラスごとに五科目の平均点を算出し、平均点の良いクラス順に並べてくださいという問題。ただし、平均点は、小数第一位に丸める必要があります。

この問題では、クラスごとあるので、まずグループ化する必要があるということ、次に平均ということで集計関数としてAVG（）を合計点をあらわす列「sum」に用います。

平均点は丸める必要がるので、平均値をROUND()関数の最初の引数に、小数第一位に

まるめるので、第二引数に１を指定します。

平均点の高い順とあるので最後にORDERBY区で降順に並べ替えることを忘れないようにしてください。

P.1１

３組の数学の点数上位5名の名前と、その点数を抽出し、点数の高い順に並べてくださいという問題。

この問題では3組の数学の点数とあるので、まずはWHERE区で項目クラスが３の生徒に絞ります。あとは問題にあるように、生徒の名前、数学の点数をセレクト句に記述します。

最後に、点数上位5名とあるので、ORDERBY区で点数が高い順に並べ替え、TOP句を使って出力を上位５行のデータに絞れば答え。

ここで、注意だがTOP句は出力されるデータから、指定した数分だけ出力する構文なので、5位で同じ点数の人が複数いたとしても、全員が抽出されるわけではないことに注意。

P.1２

理系教科（数・理）の合計が160点以上の生徒の名前と、その合計点を抽出し、  
合計点の高い順に並べてくださいという問題です。

この問題では、理系教科（数・理）の合計を表す列がもとのテーブルには存在しないので

計算する必要があります。

SQLでは項目同士、計算可能なデータ型同士、例えば数値型同士であれば

計算し、出力が可能です。

なので、今回は数学と理科の点数をあらわす項目同士を足し算し

WHERE区で問題になる160点以上の生徒に絞ります。

出力の際は、生徒の名称と含め、ORDERBYで降順にして出力することを忘れずに。

ここでは数字の足し算を実行ししたが、足し算であれば文字型の項目同士でも

実行可能で、項目同士の文字を結合した値を返すこともぜひ覚えておいてください。

実行順について。