MAT 259 > Project 1: Data Query

Grant McKenzie (grant.mckenzie@geog.ucsb.edu)

QUESTION:

One of the most interesting aspects of a dataset such as the one provided by the *Seattle Public Library*, is its ability (when asked the proper questions) to speak on user behavior . Given the completeness of the dataset, one can explore *trends* in the data and arguably infer user-interests based on these trends.

For this assignment I chose to explore how interest in subject matter changes over time. To do this I decided to investigate the variance in the items that users checked-out based on the Dewey Decimal Class of the medium. *Variance* is computed as it allows one to compare values between temporal units (e.g., month or day). A large variance indicates that media related to a wide range of classes were checked-out of the library, while a small (relative) variance indicates that the checkout history during the specified time period did not vary much in terms of subject (Dewey Class). It is important to note here that the Dewey Decimal Class system is not assigned to all media in the Library dataset. Given this, the results of the queries, as well as any inferences, are based on a subset of the library data.

Since Variance alone can often be misleading, it was important to also list the total number of items checked out as well as the count for each Dewey Class.

The two queries below show classification Variance by Month and by Day of the Week.

SQL QUERIES:

MONTH:

```
SELECT month, monthnum, VARIANCE(count) AS variance, count(*) as cnt, sum(count) as
sum
FROM
  (SELECT monthname(cin) as month, month(cin) as monthnum,
substring(deweyClass,1,3) as subdew, count(*) as count
  FROM
      (SELECT cin, deweyClass
      FROM spl2.inraw
      WHERE deweyClass <> '') as a1
  GROUP by subdew, month) as a2
GROUP BY month, monthnum
ORDER BY monthnum;
```

DAY OF WEEK:

```
SELECT day, daynum, VARIANCE(count) AS variance, count(*) as cnt, sum(count) as sum
FROM
  (SELECT dayname(cin) as day, dayofweek(cin) as daynum, substring(deweyClass,1,3)
as subdew, count(*) as count
  FROM
      (SELECT cin, deweyClass
      FROM spl2.inraw
      WHERE deweyClass <> '') as a1
  GROUP by subdew, day) as a2
GROUP BY day, daynum
ORDER BY daynum;
```

QUERY EXPLANATION:

The question I intended to answer required a two level hierarchical query (sub queries). The initial query simply returns the *Check-in Timestamp* as well as the *DeweyClass* for all records in the table *inraw* where a *DeweyClass* exists. From this query the rows are aggregated by the day of the week (using *dayname and dayofweek* functions) and the 3 digits before the decimal (class) are taken from the *DeweyClass* column using a *substring* function. Additionally, the count for all records of the specified *DeweyClass* are computed. Lastly, the day, the numerical representation of the day, the variance (of the count), the total number of unique DeweyClasses per day and the total number of records per day are reported. The above steps were repeated for month.

RESULTS:

MONTH	MONTH_NUM	VARIANCE	DDC_COUNT	TOTAL_ITEMS
January	1	625628160.2322	872	2482939
February	2	584416553.6046	886	2441234
March	3	811697936.6273	879	2918522
April	4	703141854.3939	875	2684755
May	5	666337555.9436	874	2580837
June	6	707490932.3043	875	2681675
July	7	726666001.7056	878	2711706
August	8	724558276.6155	879	2670405
September	9	496804534.169	871	2241631
October	10	604580899.4472	868	2460373
November	11	554948159.1354	879	2376709

DAY	DAY_NUM	VARIANCE	DDC_COUNT	TOTAL_ITEMS
Sunday	1	397993479.8416	874	1969910
Monday	2	2660434015.5491	897	5269587
Tuesday	3	2588466694.0293	883	5128994
Wednesday	4	2527227004.2695	893	5101534
Thursday	5	2218511032.4472	888	4747085
Friday	6	1514679123.5407	890	3929554
Saturday	7	1993389576.0715	888	4525020

PROCESSING TIME:

Month = 100.664 Seconds
Day = 106.977 Seconds

COMMENTS & ANALYSIS

What I found most interesting about this form of analysis is looking at how the variance changed over time. The *Month* query did not show any major trends that I could explain. March showed the highest variance while September showed the least amount of variance.

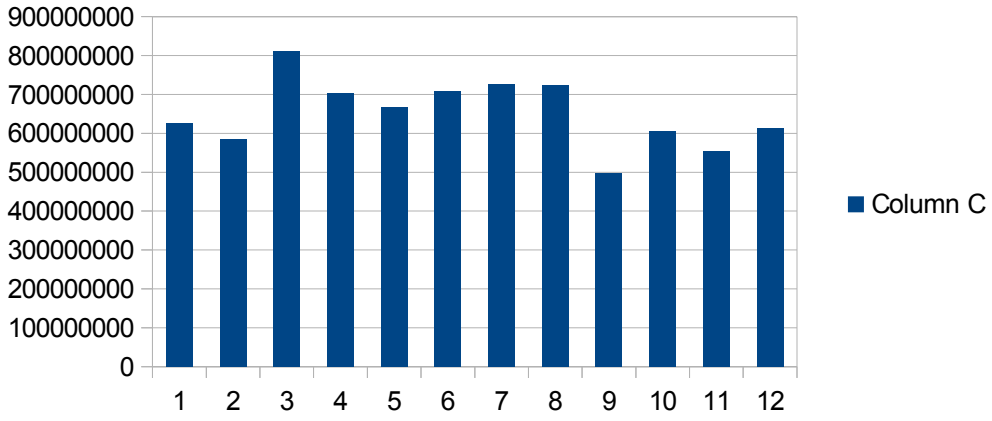

Figure 1: January – December Variance

The *Day* query on the other hand showed a trend during the week. Monday had the most variance while Friday had the least. Sunday (which has a reduced number of hours shows the least amount of variance). These results are highly correlated to the total number of items checked out.

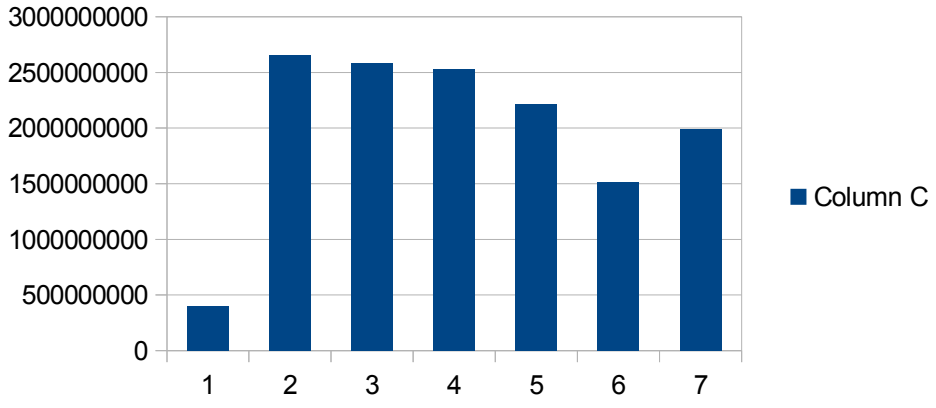

Figure 2: Sunday – Saturday Variance