Data Aggregation Activity – Module 5

Phil Roth

C0776168

It seems like the HR simulation is broken, and it will not accept my input. Here are the HDInsight cluster queries which I created, giving me the desired results. Please consider these aggregate queries as my submission for this evaluation.

Since the aggregation activity is carrying on with the transformed dataset from Assignment 4, there are a number of steps occurring before the Step #1 below. (MonthlyIncome rounded to 1000, filtered to show only the 'Sales' Dept, sorted by JobSatisfaction)

**Step #1**

CREATE TABLE sales\_attrition\_transf

(

Attrition string,

MonthlyIncome int

);

INSERT OVERWRITE TABLE sales\_attrition\_transf

SELECT Attrition, MonthlyIncome

FROM employee\_sales;

*I have named the new column as "attrition\_count", as the database already has a column named "attrition"*

SELECT COUNT(attrition) AS attrition\_count

FROM sales\_attrition\_transf

WHERE attrition LIKE "%Yes%";

CREATE TABLE sales\_non\_attrition\_transf

(

Attrition string,

MonthlyIncome int

);

INSERT OVERWRITE TABLE sales\_non\_attrition\_transf

SELECT Attrition, MonthlyIncome

FROM employee\_sales;

*I have named the new column as "non\_attrition\_count", as the database already has a column named "non\_attrition"*

SELECT COUNT(attrition) AS non\_attrition\_count

FROM sales\_non\_attrition\_transf

WHERE attrition LIKE "%No%";

**Step #2**

CREATE TABLE sales\_stats\_transf

(

Attrition string,

MonthlyIncome int

);

INSERT OVERWRITE TABLE sales\_stats\_transf

SELECT Attrition, MonthlyIncome

FROM employee\_sales;

SELECT AVG(MonthlyIncome) AS average\_monthly\_income, MIN(MonthlyIncome) AS min\_monthly\_income, MAX(MonthlyIncome) AS max\_monthly\_income

FROM sales\_stats\_transf;

SELECT AVG(MonthlyIncome) AS average\_monthly\_income, MIN(MonthlyIncome) AS min\_monthly\_income, MAX(MonthlyIncome) AS max\_monthly\_income

FROM sales\_stats\_transf

WHERE attrition LIKE "%Yes%";

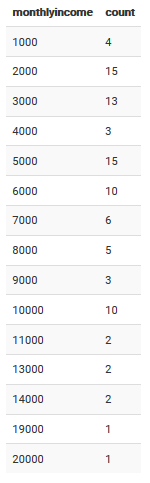
SELECT AVG(MonthlyIncome) AS average\_monthly\_income, MIN(MonthlyIncome) AS min\_monthly\_income, MAX(MonthlyIncome) AS max\_monthly\_income

FROM sales\_stats\_transf

WHERE attrition LIKE "%No%";

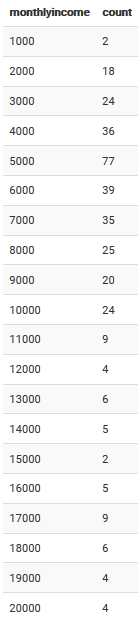
**Step #3**

SELECT MonthlyIncome, COUNT(MonthlyIncome) AS Count

FROM sales\_stats\_transf

WHERE attrition LIKE "%Yes%"

GROUP BY MonthlyIncome;

SELECT MonthlyIncome, COUNT(MonthlyIncome) AS Count

FROM sales\_stats\_transf

WHERE attrition LIKE "%No%"

GROUP BY MonthlyIncome;