**Salary Discrepancy Analysis Using SQL**

1. **Business Case:**

**As a data analyst working in a firm. The manager is seeking insights into salary disparities and wants you to identify the departments with Salaries above and below the average within each department.**

1. **Objective:**

**Create a query that identifies a high amount of variation in departments salaries.**

1. **Deliverable:**

**List from a SQL database with Average Salary, Department, and a way to score variation**

1. **Hypothesis:**

**PWD department has been flagged a department that has a high amount of salary spread.**

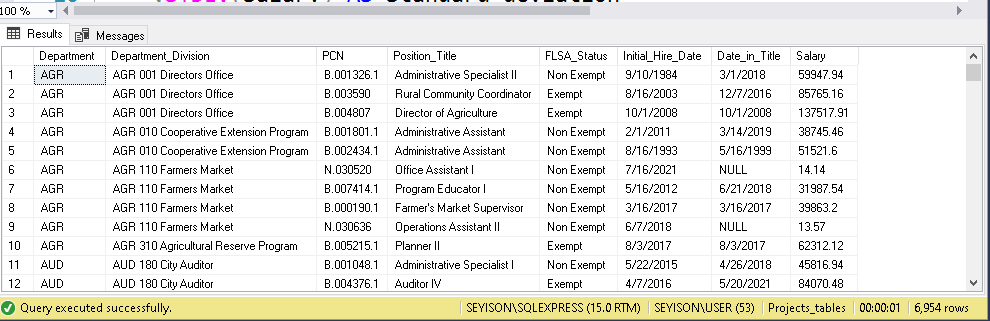
**Query Breakdown:**

* Query 1**;** Calculate the stats by departments
* Query 2; Create an outlier table
* Query 3; Filter and sort the result s by variation score.

**Analysis-in-motion:**

--query the dataset

SELECT \* FROM [dbo].[Salaries\_now]

#Result shown as:

--group by the dept and obtain the standard deviation and average

WITH Departmentstats AS (

SELECT

Department,

STDEV(Salary) AS Standard\_deviation,

AVG(Salary) AS Average

FROM [dbo].[Salaries\_now]

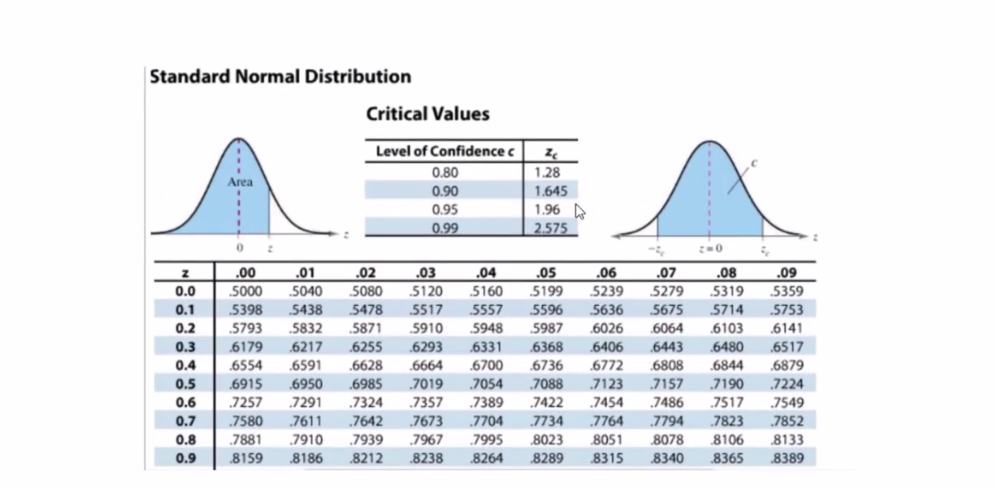
WHERE Salary >= 10000

GROUP BY Department

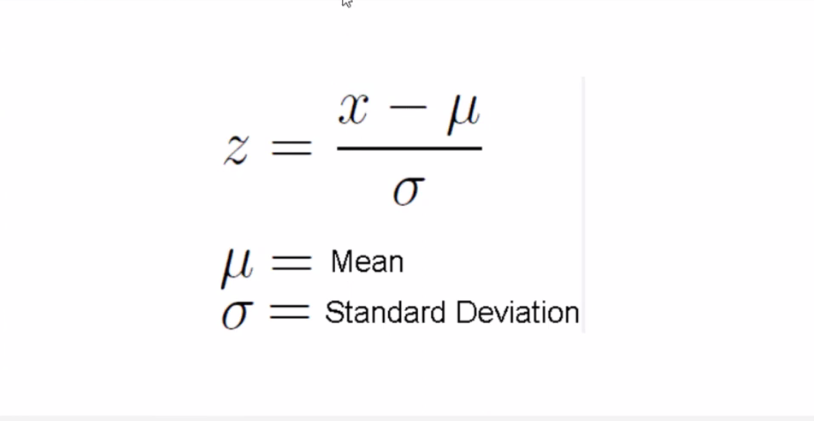
)

/\*create dept outliers

We need our zscore



To calculate our zscore



Where:

X= Salary

µ= Average

Ժ= Standard deviation

\*/

SELECT

emp.Department,

emp.salary,

dt.Standard\_Deviation,

dt.Average,

(emp.Salary-dt.Average)/dt.Standard\_Deviation AS zscore

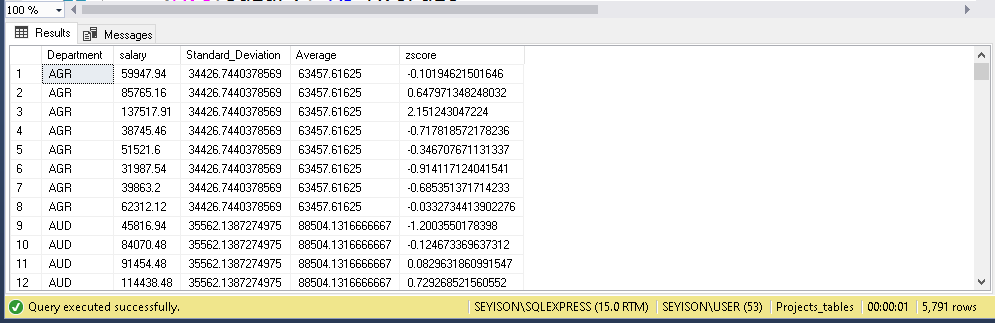
FROM [dbo].[Salaries\_now] AS emp

JOIN Departmentstats AS dt

ON emp.Department=dt.Department

WHERE emp.Salary >= 10000 ;

#Result Shown as:



-- join the two tables together using CTEs

WITH Departmentstats AS (

SELECT

Department,

STDEV(Salary) AS Standard\_deviation,

AVG(Salary) AS Average

FROM [dbo].[Salaries\_now]

WHERE Salary >= 10000

GROUP BY Department

),

DepartmentOutliers AS (

SELECT emp.Department

,emp.salary

,dt.Standard\_Deviation

,dt.Average

,(emp.Salary-dt.Average)/dt.Standard\_Deviation AS zscore

FROM [dbo].[Salaries\_now] AS emp

JOIN Departmentstats AS dt

ON emp.Department=dt.Department

WHERE emp.Salary >= 10000 )

-- get the coefficient of variation and outlier count using a CASE statement

SELECT dt.Department,

dt.Standard\_Deviation,

dt.Average,

dt.Standard\_Deviation/dt.Average AS Coefficient\_of\_Variation,

SUM(CASE WHEN (do.zscore > 1.96 OR do.zscore < -1.96 ) THEN 1 ELSE 0 END) AS Outlier\_Count

FROM Departmentstats AS dt

LEFT JOIN DepartmentOutliers AS do

ON dt.Department=do.Department

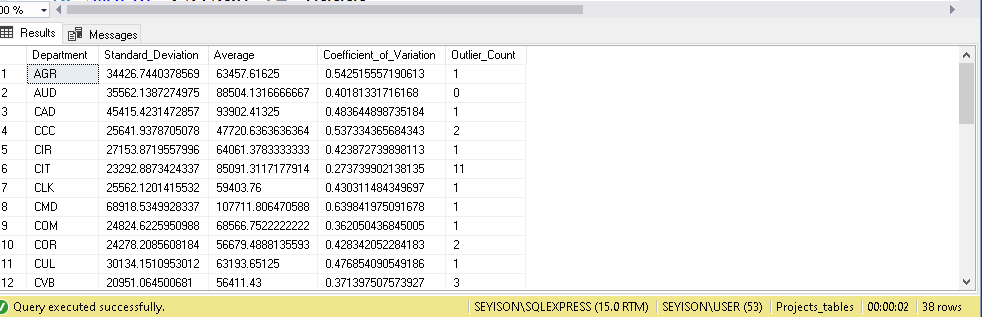
GROUP BY

dt.Department, dt.Standard\_deviation, dt.Average,

dt.Standard\_deviation/dt.Average

ORDER BY dt.Department ;

#Result shown as:



--use the round function to get rid of the trailing figures and get

WITH Departmentstats AS (

SELECT

Department,

STDEV(Salary) AS Standard\_deviation,

AVG(Salary) AS Average

FROM [dbo].[Salaries\_now]

WHERE Salary >= 10000

GROUP BY Department

),

DepartmentOutliers AS (

SELECT emp.Department,

emp.salary,

dt.Standard\_Deviation,

dt.Average,

(emp.Salary-dt.Average)/dt.Standard\_Deviation AS zscore

FROM [dbo].[Salaries\_now] AS emp

JOIN Departmentstats AS dt

ON emp.Department=dt.Department

WHERE emp.Salary >= 10000 )

SELECT dt.Department,

ROUND(dt.Standard\_Deviation,2) AS Standard\_Deviation,

ROUND(dt.Average,2) AS Salary\_Average,

ROUND((dt.Standard\_Deviation/dt.Average),2) \*100 AS Coefficient\_of\_Variation,

SUM(CASE WHEN (do.zscore > 1.96 OR do.zscore < -1.96 ) THEN 1 ELSE 0 END) AS Outlier\_Count

FROM Departmentstats AS dt

LEFT JOIN DepartmentOutliers AS do

ON dt.Department=do.Department

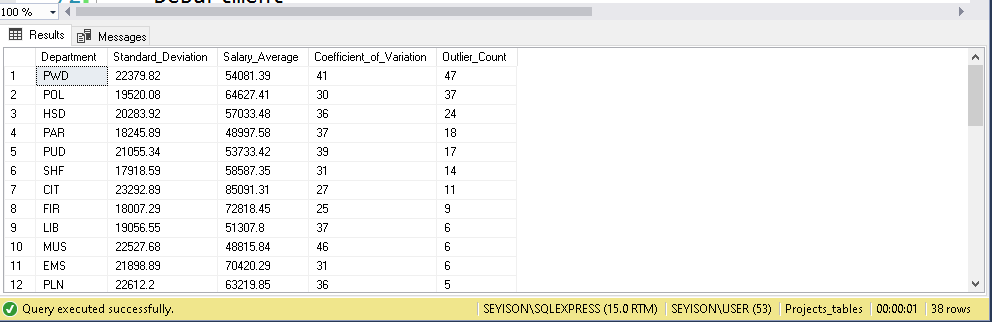
GROUP BY

dt.Department, dt.Standard\_deviation, dt.Average,

dt.Standard\_deviation/dt.Average

ORDER BY Outlier\_Count DESC ;

#Result shown as:



-- TOP 5 Departments with Salary Disparities

Based on these metrics (coefficient\_of\_Variation, outliers, Standard\_deviation) , here are the departments which shows the most variance and discrepancy in salary according to the metrics:

1. CMD: High CV(Coefficient\_of\_Variation) ,high standard deviation and some outliers
2. AGR: High CV,fairly high Standard deviation, and some outliers
3. CCC: High CV,fair number of outliers
4. PWD: High number of outliers, and fair CV
5. CWA: Fairly high standard deviation ,high CV, a fair number of outliers.