COMBINING TABLES

Combining tables

```
CREATE TABLE
`harveaspace.data.newhr_table`
AS
SELECT DISTINCT
 e.Employee_ID,
 e.First_Name,
 e.Last_Name,
 e.SSN,
 e.Birth_Date,
 e.Sex,
 e.Address,
 e.Job_ID,
 e.Salary,
 d.Department_ID,
 d.Department_Name,
 d.Location_ID,
 jh.Start_Date,
 jo.Job_Title,
 jo.Minimum_Salary,
 jo.Maximum_Salary
FROM
 `harveaspace.data.Employee` AS e
LEFT JOIN
 `harveaspace.data.departments` AS d
ON
 e.Department_ID = d.Department_ID
LEFT JOIN
 `harveaspace.data.job history` AS jh
ON
 e.Employee_ID = jh.Employee_ID
LEFT JOIN
 `harveaspace.data.jobs` AS jo
ON
 jh.Job_ID = jo.Job_ID;
```

READING DATA AND CLEANING	
Overview of table	`SELECT * FROM` `` `harveaspace.data.newhr_table` ``

Adding Rows	`INSERT INTO`
	`` `harveaspace.data.newhr_table` ``
	`(Employee_ID, First_Name, Last_Name,`
	`SSN, Birth_Date, Sex, Address, Job_ID,`
	`Salary, Department_ID, Department_Name,`
	`Location_ID, Start_Date, Job_Title,`
	`Minimum_Salary, Maximum_Salary)`
	`VALUES`
	`(1, 'Aarav', 'Sharma', '234-56-7890',`
	`'1992-11-25', 'M', '321 Pine Ave', 104,`
	`55000, 1, 'Marketing', 1, '2018-08-01',`
	`'Marketing Coordinator', 50000, 70000),`
	`(2, 'Neha', 'Patel', '876-54-3210',`
	`'1987-07-10', 'F', '654 Maple Dr', 105,`
	`60000, 1, 'Marketing', 1, '2017-12-20',`
	`'Marketing Assistant', 55000, 75000);`

Checking the Years of Experience by Employees	`SELECT`
	`Employee_ID,`
	`First_Name,`
	`Last_Name,`
	`Start_Date,`
	`DATE_DIFF(CURRENT_DATE(), Start_Date,`
	`YEAR) AS Total_Years`
	`FROM`
	`` `harveaspace.data.newhr_table` ``

Creating a new column named ”Total year”	`CREATE OR REPLACE TABLE`
	`harveaspace.data.newhr_table_new AS`
	`SELECT`
	`Employee_ID,`
	`First_Name,`
	`Last_Name,`
	`Start_Date,`

	(CASE WHEN Employee_ID IN (4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26) THEN Total_Years ELSE DATE_DIFF(CURRENT_DATE(), Start_Date, YEAR) END) AS Total_Years FROM harveaspace.data.newhr_table;
CORRECTING INCONSISTENT OR ERRONEOUS DATA	
Updated Employee Neha Patel's Address	UPDATE `harveaspace.data.newhr_table` SET Address = '123 Elm St' WHERE First_Name = 'Neha' AND Last_Name = 'Patel';
HANDLING MISSING VALUES	
Missing data checks	SELECT COUNTIF(Employee_ID IS NULL) AS Missing_Employee_ID, COUNTIF(First_Name IS NULL) AS Missing_First_Name, COUNTIF(Last_Name IS NULL) AS Missing_Last_Name, COUNTIF(SSN IS NULL) AS Missing_SSN, COUNTIF(Birth_Date IS NULL) AS Missing_Birth_Date, COUNTIF(Sex IS NULL) AS Missing_Sex, COUNTIF(Address IS NULL) AS Missing_Address, COUNTIF(Job_ID IS NULL) AS Missing_Job_ID, COUNTIF(Salary IS NULL) AS Missing_Salary, COUNTIF(Department_ID IS NULL) AS Missing_Department_ID,

	COUNTIF(Department_Name IS NULL) AS Missing_Department_Name, COUNTIF(Location_ID IS NULL) AS Missing_Location_ID, COUNTIF(Start_Date IS NULL) AS Missing_Start_Date, COUNTIF(Job_Title IS NULL) AS Missing_Job_Title, COUNTIF(Minimum_Salary IS NULL) AS Missing_Minimum_Salary, COUNTIF(Maximum_Salary IS NULL) AS Missing_Maximum_Salary FROM `harveaspace.data.newhr_table`
	-- Check if all Location_ID values exist in the Employee_ID column SELECT DISTINCT t.Location_ID FROM `harveaspace.data.newhr_table` t LEFT JOIN `harveaspace.data.newhr_table` e ON t.Location_ID = e.Employee_ID WHERE e.Employee_ID IS NULL
Manage Missing values	-- Replace missing values in Address column with 'Unknown' UPDATE `harveaspace.data.newhr_table` SET Address = 'Unknown' WHERE Address IS NULL
UNIQUE CONSTRAINT	
Duplicate checks	SELECT * FROM (SELECT *, ROW_NUMBER() OVER (PARTITION BY Employee_ID, First_Name, Last_Name, SSN, Birth_Date, Sex, Address, Job_ID, Salary, Department_ID, Department_Name,

	Location_ID, Start_Date, Job_Title, Minimum_Salary, Maximum_Salary ORDER BY Employee_ID) AS row_num FROM `harveaspace.data.newhr_table`) WHERE row_num > 1;
Duplicate checking in specific columns	-- Check for duplicate Employee_IDs SELECT Employee_ID, COUNT(*) AS Duplicate_Count FROM `harveaspace.data.newhr_table` GROUP BY Employee_ID HAVING COUNT(*) > 1
VALIDATING DATA INTEGRITY	
actual minimum and maximum values for the salary range you want to validate	SELECT * FROM `harveaspace.data.newhr_table` WHERE Salary BETWEEN 50000 AND 80000;
This statement created a new table named newhr_table_distinct.	CREATE OR REPLACE TABLE `harveaspace.data.newhr_table_distinct` AS SELECT DISTINCT * FROM `harveaspace.data.newhr_table`;
STANDARDIZING DATA FORMATS	
Department_Name length	SELECT LENGTH(Department_Name) AS Department_Name_Length FROM `harveaspace.data.newhr_table_distinct`
Find the department more than 8 letters	SELECT Department_Name FROM `harveaspace.data.newhr_table_distinct` WHERE LENGTH(Department_Name) > 8
Change “marketing” to “marketing by filed”	UPDATE `harveaspace.data.newhr_table_distinct`

	SET Department_Name = 'Marketing by Field' WHERE Department_Name = 'Marketing'
Removing spaces	-- Remove leading/trailing spaces from First_Name column UPDATE `harveaspace.data.newhr_table_distinct` SET First_Name = TRIM(First_Name) WHERE First_Name IS NOT NULL
Date Format changing	---To change the display format of the Birth_Date column from 'YYYY/MM/DD'to 'DD/MM/YYYY' without modifying the underlying data type, SELECT Employee_ID, First_Name, Last_Name, FORMAT_DATE('%d/%m/%Y', Birth_Date) AS Formatted_Birth_Date, Sex, Address, Job_ID, Salary, Department_ID, Department_Name, Location_ID, Start_Date, Job_Title, Minimum_Salary, Maximum_Salary FROM `harveaspace.data.newhr_table_distinct`
DATA TYPE VALIDATION	

Data Type Validation:	-- Check if Birth_Date is a valid date SELECT Birth_Date FROM `harveaspace.data.newhr_table_distinct` WHERE SAFE_CAST(Birth_Date AS DATE) IS NULL
HANDLING OUTLIERS	
Outlier Detection:	-- Identify outliers in Maximum Salary column using z-score SELECT * FROM (SELECT *, ABS((Maximum_Salary - AVG(Maximum_Salary) OVER ()) / STDDEV(Maximum_Salary) OVER ()) AS z_score FROM`harveaspace.data.newhr_table_distinct`) AS subquery WHERE Maximum_Salary IS NOT NULL AND z_score > 3
Removing Irrelevant Data:	-- Remove the Location ID column from the table ALTER TABLE `harveaspace.data.newhr_table_distinct` DROP COLUMN Location_ID

DATA ANALYSIS ABOUT SALARY

Questions	Execution Codes
How many employees are there in the dataset?	SELECT COUNT(*) AS Total_Employees FROM `harveaspace.data.newhr_table_distinct`

How many employees are there in each department?	SELECT Department_Name, COUNT(*) AS Total_Employees FROM `harveaspace.data.newhr_table_distinct` GROUP BY Department_Name
What is the employee count by job title?	SELECT Job_Title, COUNT(*) AS Total_Employees FROM`harveaspace.data.newhr_table_distinct` GROUP BY Job_Title
Most High Paying Departments:	SELECT Department_Name, MAX(Salary) AS Highest_Salary FROM `harveaspace.data.newhr_table_distinct` GROUP BY Department_Name ORDER BY Highest_Salary DESC;
Most Low Paying Departments:	SELECT Department_Name, MIN(Salary) AS Lowest_Salary FROM `harveaspace.data.newhr_table_distinct` GROUP BY Department_Name ORDER BY Lowest_Salary ASC;
Which are the jobs earn between 35000 to 50000?	SELECT DISTINCT Job_Title FRO`harveaspace.data.newhr_table_distinct` WHERE Salary >= 35000 AND Salary <= 50000
Most High Paying Job:	SELECT Job_Title, MAX(Salary) AS Highest_Salary FROM `harveaspace.data.newhr_table_distinct` GROUP BY Job_Title ORDER BY

	`Highest_Salary DESC;`

Most Low Paying Job:	`SELECT`
	`Job_Title,`
	`MIN(Salary) AS Lowest_Salary`
	`FROM`
	`` `harveaspace.data.newhr_table_distinct` ``
	`GROUP BY`
	`Job_Title`
	`ORDER BY`
	`Lowest_Salary ASC;`

Most High Paying 5 Employees:	`SELECT Employee_ID, First_Name, Last_Name, Salary`
	`` FROM`harveaspace.data.newhr_table_distinct` ``
	`ORDER BY Salary DESC`
	`LIMIT 5`

Most Low Paying 5 Employee:	`SELECT`
	`First_Name,`
	`Last_Name,`
	`Salary`
	`FROM`
	`` `harveaspace.data.newhr_table_distinct` ``
	`ORDER BY`
	`Salary ASC`
	`LIMIT 5;`

High Paying Person and Experience:	`SELECT`
	`First_Name,`
	`Last_Name,`
	`Salary,`
	`Start_Date,`
	`DATEDIFF(NOW(), Start_Date) AS ExperienceDays`
	`FROM`
	`` `harveaspace.data.newhr_table_distinct` ``
	`ORDER BY`
	`Salary DESC`
	`LIMIT 1;`

Low Paying Person and Experience:	SELECT
	  First_Name,
	  Last_Name,
	  Salary,
	  Start_Date,
	  DATEDIFF(NOW(), Start_Date) AS ExperienceDays
	FROM
	  `harveaspace.data.newhr_table_distinct`
	ORDER BY
	  Salary ASC
	LIMIT 1;

return the name,jobtitle,department and salary of employee that have a name of “ann”.	SELECT
	 Department_Name,
	 Salary
	FROM
	 `harveaspace.data.newhr_table_distinct`
	WHERE
	 Department_Name IN (
	  SELECT
	  Department_Name
	  FROM
	  `harveaspace.data.newhr_table_distinct`
	  WHERE
	  Firstname = “Ann”
	)

How many employees have a salary above a certain threshold (e.g., $80,000)?	SELECT COUNT(*) AS Employees_Above_Threshold FROM`harveaspace.data.newhr_table_distinct` WHERE Salary > 80000
What is the average salary by department?	SELECT Department_Name, AVG(Salary) AS Average_Salary FROM`harveaspace.data.newhr_table_distinct` GROUP BY Department_Name
Compare average,minmum and maximum salary range in department and salary	--Grouping and aggregating--- SELECT Department_Name, SUM(Salary) AS Total_Salary,

	AVG(Salary) AS Average_Salary, MIN(Salary) AS Minimum_Salary, MAX(Salary) AS Maximum_Salary FROM `harveaspace.data.newhr_table_distinct` GROUP BY Department_Name
calculates the average salary for each department by partitioning the data based on the Department_Name column.	--aggregation-- SELECT Department_Name, Salary, AVG(Salary) OVER (PARTITION BY Department_Name) AS Average_Salary FROM `harveaspace.data.newhr_table_distinct`
Sorting salary by department name	SELECT Department_Name, Salary FROM `harveaspace.data.newhr_table_distinct` ORDER BY Department_Name, Salary DESC
What is the salary range for each job title?	SELECT Job_Title, MIN(Salary) AS Minimum_Salary, MAX(Salary) AS Maximum_Salary FROM`harveaspace.data.newhr_table_distinct` GROUP BY Job_Title
What is the distribution of employees by gender?	SELECT Sex, COUNT(*) AS Total_Count FROM `harveaspace.data.newhr_table_distinct` GROUP BY Sex

ANALYSIS ABOUT RECRUITMENT

How many employees were hired in each year?	SELECT EXTRACT(YEAR FROM Start_Date) AS Hire_Year, COUNT(*) AS Total_Employees FROM`harveaspace.data.newhr_table_distinct` GROUP BY Hire_Year ORDER BY Hire_Year

ANALYSIS ABOUT COMPANY

Total payment company spends in one month?	SELECT SUM(Minimum_Salary) AS total_company_spends FROM `harveaspace.data.newhr_table_distinct`
Total payment company spends in one month for each departments?	SELECT Department_Name, SUM(Salary) AS Total_Salary FROM `harveaspace.data.newhr_table_distinct` GROUP BY Department_Name;