**Assignment: SQL to DataFrame with Cleaning, Analysis, and CSV Handling**

**Objective:**

Learn how to:

1. Create a database view combining multiple tables.
2. Load the view's data into a pandas DataFrame.
3. Clean and analyze the data.
4. Visualize insights using graphs.
5. Handle CSV export/import issues and validate data integrity.

**Assignment Steps**

**Step 1: Connect to the Database**

* Connect to a PostgreSQL database using Python.
* Write a query to list all tables in the database to understand its structure.

**Step 2: Create a Database View**

* Write a SQL query to join the following tables:
  + employees
  + departments
  + jobs
  + salary\_grades
* note! All employees must be included in the final result.
* Include the following columns in the view:
  + employee\_id, first\_name, last\_name, salary, commission\_pct
  + department\_id, department\_name
  + job\_id, job\_title
  + grade\_level (based on salary range from salary\_grades)
* Save the query as a view named employee\_details\_view.

**Step 3: Load Data into pandas**

* Query the employee\_details\_view and load the result into a pandas DataFrame.

**Step 4: Data Quality Checks**

* Check for missing values in the dataset.
* Set emp\_id to be the index
* Display summary statistics (e.g., mean, median, min, max) for numerical columns.
* Identify any potential outliers in the salary column.

**Step 5: Clean the Data**

1. Replace NULL values in:
   * commission\_pct with 0
   * department\_name with "Unknown"
2. Drop the department id and job id columns
3. Identify and change the appropriate columns to category type

**Step 6: Data Analysis and Visualization**

1. Create a **bar chart** showing the average salary by department.
2. Create a **pie chart** showing the number of employees by salary grade.
3. Create a **bar chart** showing the number of employees by job title.
4. Create a **heatmap** showing correlations between numerical columns like salary and commission\_pct.

**Step 7: Export Data to CSV**

1. Export the cleaned DataFrame to a CSV file.
2. Reload the CSV file into pandas and check for an extra index column.
3. If the extra index column exists:
   * Remove it.
   * Re-export the corrected file.

**Step 8: Verify Data Integrity**

* Compare the original DataFrame with the reloaded DataFrame to ensure no important data was lost during the CSV export/import process.

**Deliverables**

**(לא באמת צריך להגיש, אבל תעבדו מסודר ותשמרו לכם יחד את הקוד, את הcsv, וגם כל גרף כקובץ תמונה -בתיקיה מאושרת ושמחה אחת)**

1. A Python script or Jupyter Notebook containing:
   * The SQL query for the view.
   * Code for loading, cleaning, analyzing, and visualizing the data.
   * Code for handling the CSV export/import and validating data integrity.
2. Two CSV files:
   * cleaned\_employees.csv (with potential extra index column).
   * corrected\_employees.csv (without extra index column).
3. Graphs:
   * Bar chart of average salary by department.
   * Pie chart of employees by salary grade.
   * Bar chart of employees by job title.
   * Heatmap of numerical correlations.