**BSD 3203 Programming for Data Science**

**CAT II based on Practical in Programming for Data Science (R & Python)**

**Duration:** 2 Hours  
**Total Marks:** 40  
**Instructions:**

* Attempt **both** sections (R and Python), INDIVIDUALLY. You are required to answer **all tasks** in both **R and Python**.
* Install your preferred softwares (VS code, Spyder, Jupiter Notebook, Pytorch, R&R Studio) for simulating your answers. Show **screenshots** of your code and outputs where applicable. The screenshots should capture your entire PC screen including date and time. Compile one submission and name as BSD3203\_CAT2\_YourName.R\_py.
* Ensure your code is **well commented** for clarity.
* Submit both your **R script (.R)** and **Python script (.py)** files.
* Use the provided datasets to complete the tasks.
* CAT Summary and Distribution of Marks

|  |  |  |
| --- | --- | --- |
| **Section** | **Question** | **Marks** |
| **Section A (R)** | Data Manipulation (Q1) | 6 |
|  | Data Visualization (Q2) | 6 |
|  | Statistical Analysis (Q3) | 8 |
| **Section B (Python)** | Data Manipulation (Q4) | 6 |
|  | Data Visualization (Q5) | 6 |
|  | Machine Learning - Linear Regression (Q6) | 8 |
| **Total Marks** |  | **40** |

**DATASETS**

Create these CSV files for use in the tasks by copying the tables into excel and saving as CSV files:

1. sales\_data.csv
2. employee\_data.csv
3. house\_prices.csv (for Python ML section)

**Sample Datasets (CSV Format)**

**1. sales\_data.csv (for R section)**

|  |  |  |  |
| --- | --- | --- | --- |
| Product | Region | Sales | Date |
| Laptop | Region A | 7000 | 1/10/2024 |
| Phone | Region B | 4500 | 1/12/2024 |
| Tablet | Region A | 8000 | 1/15/2024 |
| Laptop | Region B | 3000 | 1/18/2024 |
| Phone | Region A | 6500 | 1/20/2024 |

**2. employee\_data.csv (for Python section)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Department** | **Salary** | **Hire\_Date** |
| John Doe | HR | 55000 | 6/15/2020 |
| Jane Smith | IT | 75000 | 3/22/2018 |
| Mike Brown | Sales | 48000 | 10/10/2019 |
| Anna White | IT | 82000 | 1/5/2021 |
| Emma Davis | Sales | 61000 | 7/19/2017 |

**3. house\_prices.csv (for Python ML section)**

|  |  |  |
| --- | --- | --- |
| **Square\_Feet** | **Bedrooms** | **Price** |
| 1500 | 3 | 300000 |
| 1800 | 4 | 350000 |
| 2400 | 4 | 450000 |
| 3000 | 5 | 600000 |
| 1200 | 2 | 250000 |

**Section A: R Programming (20 Marks)**

**Question 1: Data Manipulation with dplyr (6 Marks)**

You are given a CSV file called sales\_data.csv with columns: Product, Region, Sales, and Date.  
Tasks:  
a) Load the dataset into R. (1 Mark)  
b) Display only the Product and Sales columns. (1 Mark)  
c) Filter rows where Sales is greater than 5000. (1 Mark)  
d) Group the data by Region and calculate the total sales per region. (2 Marks)  
e) Save the summarized data to a new CSV file called region\_sales.csv. (1 Mark)

**Question 2: Data Visualization with ggplot2 (6 Marks)**

Using the same sales\_data.csv:  
a) Create a bar chart showing total sales for each Product. (3 Marks)  
b) Create a line plot showing sales trends over time for any one region of your choice. (3 Marks)

**Question 3: Statistical Analysis (8 Marks)**

a) Perform a summary() of the Sales column. (2 Marks)  
b) Test whether the mean sales for Region A are significantly different from Region B using a t-test. (4 Marks)  
c) Interpret the p-value in your t-test result. (2 Marks)

**Section B: Python Programming (20 Marks)**

**Question 4: Data Manipulation with Pandas (6 Marks)**

You are given a CSV file called employee\_data.csv with columns: Name, Department, Salary, and Hire\_Date.  
Tasks:  
a) Load the dataset into Python (Pandas). (1 Mark)  
b) Display only the Name and Salary columns. (1 Mark)  
c) Filter records where Salary is above 60000. (1 Mark)  
d) Group the data by Department and calculate the average salary per department. (2 Marks)  
e) Save the summarized data to department\_salary.csv. (1 Mark)

**Question 5: Data Visualization with Matplotlib / Seaborn (6 Marks)**

Using the same employee\_data.csv:  
a) Create a histogram of the Salary distribution. (3 Marks)  
b) Create a boxplot showing Salary by Department. (3 Marks)

**Question 6: Machine Learning - Linear Regression (8 Marks)**

You are given a dataset house\_prices.csv with columns: Square\_Feet, Bedrooms, and Price.  
Tasks:  
a) Load the dataset into Python. (1 Mark)  
b) Use scikit-learn to create a Linear Regression model that predicts Price based on Square\_Feet and Bedrooms. (3 Marks)  
c) Display the model coefficients and intercept. (2 Marks)  
d) Predict the price of a house with 2000 square feet and 3 bedrooms. (2 Marks)

**END OF CAT**