PySpark RDD – Student Worksheet (30 Descriptive Problems)

1. Total quantity sold per product Input: [('P1',2),('P2',3),('P1',5),('P3',1),('P2',4),('P3',3),('P4',10),('P1',1)] Task: Calculate the total quantity sold for every product.

Answer: _____

2. Top 3 most viewed products Input: ['P1','P2','P1','P3','P2','P1','P4','P5','P5','P3','P5','P1'] Task: Return the top 3 most viewed items.

Answer: _____

3. Market Basket item pairs Input: [('O1',['P1','P2','P3']),('O2',['P2','P3']),('O3',['P1','P3','P4']),('O4',['P2','P4']),('O5',['P1','P2'])] Task: Generate item pairs bought together and count frequency.

Answer: _____

4. City-level sales summary Input: [('Hyd',2000),('Hyd',1500),('Blr',3000),('Hyd',500),('Chennai',2200),('Blr',1000),('Delhi',3500),('Delhi',1500)] Task: Compute total sales per city.

Answer: _____

5. Customers with more than 3 purchases Input: [('u1','P1'),('u1','P2'),('u2','P1'),('u1','P3'),('u3','P1'),('u1','P4'),('u2','P3'),('u2','P5'),('u2','P2')] Task: Identify customers with more than 3 purchases.

Answer: _____

6. Total transaction amount per user Input: [('u1',200),('u1',500),('u2',300),('u3',800),('u2',700),('u3',150),('u1',100),('u4',1000),('u4',200)] Task: Sum all transactions per user.

Answer: _____

7. High-value transactions > 10000 Input: [500,12000,220,15000,9999,18000,25000,600,45000,13000] Task: Extract transactions > 10,000.

Answer: _____

8. Credit score bucket counts Input: [720,690,810,550,620,780,640,850,590,705,765,830] Task: Categorize into Excellent, Good, Fair, Bad.

Answer: _____

9. Drug prescription frequency Input:
['Dolo','Aspirin','Dolo','Zifi','Aspirin','Dolo','Metformin','Zifi','Atorvastatin','Dolo'] Task: Count total prescriptions per drug.

Answer: _____

10. Patient visit frequency Input: [('p1',1),('p1',1),('p2',1),('p3',1),('p2',1),('p1',1),('p4',1)] Task: Count visits per patient.

Answer: _____

11. Top 3 prescribed medicines Input:
['Atorvastatin','Metformin','Dolo','Dolo','Zifi','Dolo','Metformin','Atorvastatin','Dolo','Zifi','Zifi'] Task: Return the top 3 medicines by prescription count.

Answer: _____

12. Drug pair frequency Input: [('PR1',['Dolo','Zifi','Aspirin']),('PR2',['Zifi','Metformin']),('PR3',['Dolo','Aspirin']),('PR4',['Dolo','Atorvastatin']),('PR5',['Zifi','Aspirin'])] Task: Generate all drug pairs and count their occurrences.

Answer: _____

13. High engine temperatures Input: [45,78,120,140,90,200,55,160,180,40,130,95,210] Task: Detect high engine temperatures above 120.

Answer: _____

14. Complaints per car model Input: [('Nexon','brakes'),('Punch','lights'),('Nexon','engine'),('Nexon','ac'),('Punch','engine'),('Harrier','ac'),('Nexon','suspension'),('Harrier','brakes')] Task: Count the number of complaints per car model.

Answer: _____

15. Service cost per job card Input:
[('job1',500),('job1',300),('job2',900),('job1',200),('job3',700),('job2',300),('job3',150)] Task: Compute total cost per job.

Answer: _____

16. Total call minutes per user Input:
[('u1',5),('u1',10),('u2',50),('u3',20),('u2',10),('u3',15),('u1',40),('u4',30)] Task: Sum the call duration per user.

Answer: _____

17. Identify spam callers Input:
[('n1',5),('n1',20),('n1',10),('n1',12),('n2',3),('n2',1),('n3',40),('n3',5),('n4',10)] Task: Find callers with 40+ total calls.

Answer: _____

18. Data usage per telecom circle Input:
[('AP',2.5),('KA',1.2),('AP',3.5),('KA',2.3),('TN',4.1),('TN',1.4),('AP',1.5)] Task: Calculate total data usage per circle.

Answer: _____

19. Average delivery time per city Input:
[('Hyd',2),('Hyd',3),('Delhi',5),('Blr',4),('Hyd',1),('Blr',3),('Delhi',4),('Chennai',6)] Task: Compute the average delivery time per city.

Answer: _____

20. Deliveries per driver Input: [('d1',1),('d2',1),('d1',1),('d3',1),('d2',1),('d1',1),('d3',1),('d3',1),('d4',1)] Task: Count deliveries per driver.

Answer: _____

21. Attendance per student Input: [('s1','class1'),('s1','class2'),('s1','class3'),('s2','class1'),('s3','class1'),('s1','class4'),('s2','class2'),('s2','class3')] Task: Count classes attended by each student.

Answer: _____

22. Word frequency in feedback Input: ['Good content and delivery','More projects needed','Good trainer and good content','Need more case studies','Projects case studies and notes'] Task: Count word frequency.

Answer: _____

23. Assignment submission rate Input: [('s1',1),('s1',0),('s2',1),('s3',1),('s3',0),('s2',0),('s1',1)] Task: Count submissions per student.

Answer: _____

24. Unique IP count Input:
['10.0.0.1','10.0.0.2','10.0.0.1','10.0.0.3','10.0.0.2','10.0.0.4','10.0.0.5','10.0.0.3'] Task: Count unique IP addresses.

Answer: _____

25. Distinct union of product lists Input: ['A','B','C','D'] and ['C','D','E','F'] Task: Perform union and remove duplicates.

Answer: _____

26. Common users between two platforms Input: ['u1','u2','u3','u4'] and ['u3','u4','u5','u6'] Task: Find intersection of users.

Answer: _____

27. Salary hike Input: [('e1',50000),('e2',60000),('e3',45000)] Task: Apply 10% salary hike.

Answer: _____

28. Clean invalid records Input: ['ram','',None,'shyam',' ','john'] Task: Remove missing or empty values.

Answer: _____

29. Employee count per location Input: [('Hyd','E1'),('Hyd','E2'),('Blr','E3'),('Pune','E4'),('Hyd','E5'),('Blr','E6')] Task: Count employees per location.

Answer: _____

30. Join employee and department budgets Input: employees=[('e1','HR'),('e2','IT'),('e3','Finance'),('e4','HR')], departments=[('HR','100'),('Finance','200'),('IT','150')] Task: Perform join to attach department budgets.

Answer: _____