20 Real-Life Datasets with Problem Statements & Solutions

1. Dataset: Sales Dataset

Problem: Calculate total revenue per region

Solution: df.groupby('Region')['Revenue'].sum()

2. Dataset: Sales Dataset

Problem: Identify the product with the highest total sales volume

Solution: df.groupby('Product')['Units_Sold'].sum().idxmax()

3. Dataset: COVID-19 Dataset

Problem: Compute total confirmed cases per continent

Solution: df.groupby('Continent')['Confirmed'].sum()

4. Dataset: COVID-19 Dataset

Problem: Find the day with the highest global death toll

Solution: df.groupby('Date')['Deaths'].sum().idxmax()

5. Dataset: FIFA Dataset

Problem: Find average player ratings by nationality

Solution: df.groupby('Nationality')['Overall'].mean().sort_values(ascending=False)

6. Dataset: FIFA Dataset

Problem: Get the correlation between 'Age' and 'Potential'

Solution: df[['Age', 'Potential']].corr()

7. Dataset: Cricket World Cup Dataset

Problem: Total runs scored by each team in the tournament

Solution: df.groupby('Team')['Runs'].sum()

8. Dataset: Cricket World Cup Dataset

Problem: Top 5 highest wicket-takers

Solution: df.groupby('Bowler')['Wickets'].sum().sort_values(ascending=False).head(5)

9. Dataset: IPL

Problem: Number of wins by each team

Solution: df['Winner'].value_counts()

10. Dataset: IPL

Problem: Most valuable player based on total runs and strike rate

Solution: df.groupby('Player')[['Runs', 'StrikeRate']].mean().sort_values(by='Runs', ascending=False).head(1)

11. Dataset: Kaggle Text Classification Dataset

Problem: Count of documents per category

Solution: df['label'].value_counts()

12. Dataset: Kaggle Text Classification Dataset

Problem: Average word count per document

Solution: df['text'].apply(lambda x: len(x.split())).mean()

13. Dataset: Movie Review

Problem: Proportion of positive vs. negative reviews

Solution: df['sentiment'].value_counts(normalize=True)

14. Dataset: Movie Review

Problem: Most frequent word in negative reviews

Solution: from collections import Counter

negative_words = ' '.join(df[df['sentiment'] == 'negative']['review']).split()

Counter(negative_words).most_common(1)

15. Dataset: OpinRank Review Dataset

Problem: Average rating of each product

Solution: df.groupby('ProductID')['Rating'].mean()

16. Dataset: OpinRank Review Dataset

Problem: Most mentioned feature in reviews

Solution: from collections import Counter

feature_words = ' '.join(df['ReviewText']).lower().split()

Counter(feature_words).most_common(1)

17. Dataset: Amazon Product Dataset

Problem: Count of reviews per product category

Solution: df['Category'].value_counts()

18. Dataset: Amazon Product Dataset

Problem: Average star rating per product

Solution: df.groupby('ProductID')['StarRating'].mean()

19. Dataset: Paper Review

Problem: Average reviewer score by area of research

Solution: df.groupby('Area')['Score'].mean()

20. Dataset: Paper Review

Problem: Count of accepted vs. rejected papers

Solution: df['Decision'].value_counts()