Data-Driven Construction Project Performance Analytics: Tracking Efficiency, Timelines, and Risks

Step 1: Load and Inspect Your Data

In [2]:

```
import pandas as pd

# Load the two CSV files
forms_df = pd.read_csv("Construction_Data_PM_Forms_All_Projects.csv")
tasks_df = pd.read_csv("Construction_Data_PM_Tasks_All_Projects.csv")
```

In [16]:

```
# Check the first few rows
print("Forms Data:")
print(forms_df.head())
```

```
Forms Data:
         Ref                Status  \
0  F145185.4                Opened   
1  F1.495500  Open / Ongoing Works   
2  F1.495499  Open / Ongoing Works   
3  F1.495498  Open / Ongoing Works   
4  F1.495496  Open / Ongoing Works   

                                            Location  \
0  01 Daily Site Diary>Site Management>JPC Projec...   
1  02 Daily Work Plan>Site Management>JPC Project...   
2  02 Daily Work Plan>Site Management>JPC Project...   
3  02 Daily Work Plan>Site Management>JPC Project...   
4  02 Daily Work Plan>Site Management>JPC Project...   

                             Name     Created                       Type  \
0   1328 CM-SM-FRM-001 Site Diary  15/09/2020            Site Management   
1  SM-FRM-SUB-101 Daily Work Plan  15/09/2020  Subcontractor Inspections   
2  SM-FRM-SUB-101 Daily Work Plan  15/09/2020  Subcontractor Inspections   
3  SM-FRM-SUB-101 Daily Work Plan  15/09/2020  Subcontractor Inspections   
4  SM-FRM-SUB-101 Daily Work Plan  15/09/2020  Subcontractor Inspections   

  Status Changed  Open Actions  Total Actions Association  OverDue  Images  \
0     15/09/2020             0              0         NaN    False    True   
1     15/09/2020             0              0         NaN    False   False   
2     15/09/2020             0              0         NaN    False   False   
3     15/09/2020             0              0         NaN    False   False   
4     15/09/2020             0              0         NaN    False   False   

   Comments Documents  Project Report Forms Status Report Forms Group  
0     False     False     1328                Open    Site Management  
1     False     False     1328                Open      Subcontractor  
2     False     False     1328                Open      Subcontractor  
3     False     False     1328                Open      Subcontractor  
4     False     False     1328                Open      Subcontractor  
```

In [20]:

```python
print("\nTasks Data:")
print(tasks_df.head())
```

```
Tasks Data:
           ref                status  \
0  T1.23963030                  Open   
1  T116412.200                Closed   
2   T141663.27  EHS Good Observation   
3  T116412.199                Closed   
4   T141663.26  EHS Good Observation   

                                            location  \
0  JPC Project Management>EHS Management>01 Inspe...   
1  QC & BC(A)R>ITP 02 Architectural & M&E Service...   
2  JPC Project Management>EHS Management>01 Inspe...   
3  QC & BC(A)R>ITP 02 Architectural & M&E Service...   
4  JPC Project Management>EHS Management>01 Inspe...   

                                         description     created  target  \
0  task raised in incorrect location of this form...  14/09/2020     NaN   
1                                             Metsec  14/09/2020     NaN   
2  Good clear exclusion zones and access through ...  14/09/2020     NaN   
3                                           RC walls  14/09/2020     NaN   
4  block 02 working level has good housekeeping, ...  14/09/2020     NaN   

                                       type             to package  \
0     Safety Notice (Amber) - General Issue        Main Contractor   
1                      JPC - Progress Photo  Ceilings & Partitions   
2  Safety Notice (Green) - Good Observation        Main Contractor   
3                      JPC - Progress Photo       Precast Concrete   
4  Safety Notice (Green) - Good Observation       Precast Concrete   

  status changed  association overdue images comments documents  \
0     14/09/2020  FormAnswer    False    NaN      NaN       NaN   
1     14/09/2020         NaN    False   True    False     False   
2     14/09/2020  FormAnswer    False   True    False     False   
3     14/09/2020         NaN    False   True    False     False   
4     14/09/2020  FormAnswer    False   True    False     False   

              priority                    cause  project report status  \
0  Behavioural Failure  JPC - Safety - Documentation     1328          Open   
1                  NaN                           NaN     1328        Closed   
2                  NaN         JPC - Safety - Access     1328        Closed   
3                  NaN                           NaN     1328        Closed   
4                  NaN  JPC - Safety - House Keeping     1328        Closed   

        task group  
0           Safety  
1  Site Management  
2           Safety  
3  Site Management  
4           Safety  
```

In [7]:

```
# Check basic info
print("\nForms Data Info:")
forms_df.info()
```

```
Forms Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10254 entries, 0 to 10253
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   Ref                  10254 non-null  object
 1   Status               10254 non-null  object
 2   Location             10254 non-null  object
 3   Name                 10254 non-null  object
 4   Created              10254 non-null  object
 5   Type                 10254 non-null  object
 6   Status Changed       10254 non-null  object
 7   Open Actions         10254 non-null  int64 
 8   Total Actions        10254 non-null  int64 
 9   Association          2098 non-null   object
 10  OverDue              10254 non-null  bool  
 11  Images               10254 non-null  bool  
 12  Comments             10254 non-null  bool  
 13  Documents            9450 non-null   object
 14  Project              10254 non-null  int64 
 15  Report Forms Status  10252 non-null  object
 16  Report Forms Group   10250 non-null  object
dtypes: bool(3), int64(3), object(11)
memory usage: 1.1+ MB
```

In [8]:

```python
print("\nTasks Data Info:")
tasks_df.info()
```

```
Tasks Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12424 entries, 0 to 12423
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Ref             12424 non-null  object 
 1   Status          12424 non-null  object 
 2   Location        12424 non-null  object 
 3   Description     12424 non-null  object 
 4   Created         12424 non-null  object 
 5   Target          2568 non-null   float64
 6   Type            12424 non-null  object 
 7   To Package      11382 non-null  object 
 8   Status Changed  12424 non-null  object 
 9   Association     9483 non-null   object 
 10  OverDue         12424 non-null  bool   
 11  Images          12272 non-null  object 
 12  Comments        11902 non-null  object 
 13  Documents       11780 non-null  object 
 14  Priority        2366 non-null   object 
 15  Cause           9683 non-null   object 
 16  project         12424 non-null  int64  
 17  Report Status   12424 non-null  object 
 18  Task Group      12374 non-null  object 
dtypes: bool(1), float64(1), int64(1), object(16)
memory usage: 1.7+ MB
```

In [9]:

```
# Check missing values
print("\nMissing values in Forms Data:")
print(forms_df.isnull().sum())
```

```
Missing values in Forms Data:
Ref                       0
Status                    0
Location                  0
Name                      0
Created                   0
Type                      0
Status Changed            0
Open Actions              0
Total Actions             0
Association            8156
OverDue                   0
Images                    0
Comments                  0
Documents               804
Project                   0
Report Forms Status       2
Report Forms Group        4
dtype: int64
```

In [10]:

```
print("\nMissing values in Tasks Data:")
print(tasks_df.isnull().sum())
```

```
Missing values in Tasks Data:
Ref                   0
Status                0
Location              0
Description           0
Created               0
Target             9856
Type                  0
To Package         1042
Status Changed        0
Association        2941
OverDue               0
Images              152
Comments            522
Documents           644
Priority          10058
Cause              2741
project               0
Report Status         0
Task Group           50
dtype: int64
```

Step 2: Understand Column Meanings & Data Types

In [11]:

```
# Check column names
print("Forms Columns:", forms_df.columns.tolist())
```

```
Forms Columns: ['Ref', 'Status', 'Location', 'Name', 'Created', 'Type', 'Status Chan
ged', 'Open Actions', 'Total Actions', 'Association', 'OverDue', 'Images', 'Comment
s', 'Documents', 'Project', 'Report Forms Status', 'Report Forms Group']
```

In [12]:

```
print("Tasks Columns:", tasks_df.columns.tolist())
```

```
Tasks Columns: ['Ref', 'Status', 'Location', 'Description', 'Created', 'Target', 'Ty
pe', 'To Package', 'Status Changed', 'Association', 'OverDue', 'Images', 'Comments',
'Documents', 'Priority', 'Cause', 'project', 'Report Status', 'Task Group']
```

In [13]:

```
# Get quick stats for numeric columns
print("\nForms Data Statistics:")
print(forms_df.describe())
```

```
Forms Data Statistics:
       Open Actions  Total Actions       Project
count   10254.00000   10254.000000  10254.000000
mean        0.05315       0.864541   1331.690072
std         0.54720       2.702933      5.143594
min         0.00000       0.000000   1328.000000
25%         0.00000       0.000000   1328.000000
50%         0.00000       0.000000   1329.000000
75%         0.00000       0.000000   1335.000000
max        19.00000      31.000000   1345.000000
```

In [14]:

```
print("\nTasks Data Statistics:")
print(tasks_df.describe())
```

```
Tasks Data Statistics:
             Target       project
count   2568.000000  12424.000000
mean   43968.516355   1332.585480
std      105.261518      5.213831
min    43590.000000   1328.000000
25%    43923.750000   1328.000000
50%    44000.000000   1330.000000
75%    44041.000000   1338.000000
max    44106.000000   1345.000000
```

Step 3: Merge the Datasets

In [19]:

```
# Make all column names lowercase to avoid case mismatches
forms_df.columns = forms_df.columns.str.lower()
tasks_df.columns = tasks_df.columns.str.lower()

# Merge on 'project'
merged_df = pd.merge(forms_df, tasks_df, on="project", how="inner", suffixes=("_for

print("Merged Shape:", merged_df.shape)
print(merged_df.head())
```

```
Merged Shape: (26460209, 35)
    ref_form status_form                                      location_form  \
0  F145185.4      Opened  01 Daily Site Diary>Site Management>JPC Projec...   
1  F145185.4      Opened  01 Daily Site Diary>Site Management>JPC Projec...   
2  F145185.4      Opened  01 Daily Site Diary>Site Management>JPC Projec...   
3  F145185.4      Opened  01 Daily Site Diary>Site Management>JPC Projec...   
4  F145185.4      Opened  01 Daily Site Diary>Site Management>JPC Projec...   

                          name created_form        type_form  \
0  1328 CM-SM-FRM-001 Site Diary   15/09/2020  Site Management   
1  1328 CM-SM-FRM-001 Site Diary   15/09/2020  Site Management   
2  1328 CM-SM-FRM-001 Site Diary   15/09/2020  Site Management   
3  1328 CM-SM-FRM-001 Site Diary   15/09/2020  Site Management   
4  1328 CM-SM-FRM-001 Site Diary   15/09/2020  Site Management   

  status changed_form  open actions  total actions association_form  ...  \
0          15/09/2020             0              0              NaN  ...   
1          15/09/2020             0              0              NaN  ...   
2          15/09/2020             0              0              NaN  ...   
3          15/09/2020             0              0              NaN  ...   
4          15/09/2020             0              0              NaN  ...   

   status changed_task  association_task  overdue_task images_task  \
0           14/09/2020        FormAnswer         False         NaN   
1           14/09/2020               NaN         False        True   
2           14/09/2020        FormAnswer         False        True   
3           14/09/2020               NaN         False        True   
4           14/09/2020        FormAnswer         False        True   

   comments_task documents_task             priority  \
0            NaN            NaN  Behavioural Failure   
1          False          False                  NaN   
2          False          False                  NaN   
3          False          False                  NaN   
4          False          False                  NaN   

                          cause report status       task group  
0  JPC - Safety - Documentation          Open           Safety  
1                           NaN        Closed  Site Management  
2         JPC - Safety - Access        Closed           Safety  
3                           NaN        Closed  Site Management  
4  JPC - Safety - House Keeping        Closed           Safety  

[5 rows x 35 columns]
```

In [21]:

```
# Check for duplicates in merged data
duplicates_count = merged_df.duplicated().sum()
print(f"Duplicates in merged data: {duplicates_count}")
```

```
Duplicates in merged data: 0
```

In [22]:

```
# Check column names
print("merged Columns:", merged_df.columns.tolist())
```

```
merged Columns: ['ref_form', 'status_form', 'location_form', 'name', 'created_form',
'type_form', 'status changed_form', 'open actions', 'total actions', 'association_fo
rm', 'overdue_form', 'images_form', 'comments_form', 'documents_form', 'project', 'r
eport forms status', 'report forms group', 'ref_task', 'status_task', 'location_tas
k', 'description', 'created_task', 'target', 'type_task', 'to package', 'status chan
ged_task', 'association_task', 'overdue_task', 'images_task', 'comments_task', 'docu
ments_task', 'priority', 'cause', 'report status', 'task group']
```

In [23]:

```
# standardize column names (lowercase, no spaces)
merged_df.columns = merged_df.columns.str.strip().str.lower().str.replace(' ', '_')
```

In [24]:

```
#Quick inspect (shape, types, sample)
print("shape:", merged_df.shape)
print("\ncolumns:", merged_df.columns.tolist())
print("\ninfo:")
print(merged_df.info())
print("\nhead:")
print(merged_df.head().T)   # transpose for easy view
```

```
shape: (26460209, 35)

columns: ['ref_form', 'status_form', 'location_form', 'name', 'created_form', 'type_
form', 'status_changed_form', 'open_actions', 'total_actions', 'association_form',
'overdue_form', 'images_form', 'comments_form', 'documents_form', 'project', 'report
_forms_status', 'report_forms_group', 'ref_task', 'status_task', 'location_task', 'd
escription', 'created_task', 'target', 'type_task', 'to_package', 'status_changed_ta
sk', 'association_task', 'overdue_task', 'images_task', 'comments_task', 'documents_
task', 'priority', 'cause', 'report_status', 'task_group']

info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26460209 entries, 0 to 26460208
Data columns (total 35 columns):
 #   Column               Dtype  
---  ------               -----  
 0   ref_form             object 
 1   status_form          object 
 2   location_form        object 
 3   name                 object 
 4   created_form         object 
 5   type_form            object 
 6   status_changed_form  object 
 7   open_actions         int64  
 8   total_actions        int64  
 9   association_form     object 
 10  overdue_form         bool   
 11  images_form          bool   
 12  comments_form        bool   
 13  documents_form       object 
 14  project              int64  
 15  report_forms_status  object 
 16  report_forms_group   object 
 17  ref_task             object 
 18  status_task          object 
 19  location_task        object 
 20  description          object 
 21  created_task         object 
 22  target               float64
 23  type_task            object 
 24  to_package           object 
 25  status_changed_task  object 
 26  association_task     object 
 27  overdue_task         bool   
 28  images_task          object 
 29  comments_task        object 
 30  documents_task       object 
 31  priority             object 
 32  cause                object 
 33  report_status        object 
 34  task_group           object 
dtypes: bool(4), float64(1), int64(3), object(27)
memory usage: 6.2+ GB
None

head:
```

```
                                                                   0  \
ref_form                                                 F145185.4
status_form                                                 Opened
location_form        01 Daily Site Diary>Site Management>JPC Projec...
name                                   1328 CM-SM-FRM-001 Site Diary
created_form                                              15/09/2020
type_form                                            Site Management
status_changed_form                                       15/09/2020
open_actions                                                       0
total_actions                                                      0
association_form                                                 NaN
overdue_form                                                   False
images_form                                                     True
comments_form                                                  False
documents_form                                                 False
project                                                         1328
report_forms_status                                             Open
report_forms_group                                   Site Management
ref_task                                                 T1.23963030
status_task                                                     Open
location_task        JPC Project Management>EHS Management>01 Inspe...
description          task raised in incorrect location of this form...
created_task                                              14/09/2020
target                                                           NaN
type_task                       Safety Notice (Amber) - General Issue
to_package                                           Main Contractor
status_changed_task                                       14/09/2020
association_task                                          FormAnswer
overdue_task                                                   False
images_task                                                      NaN
comments_task                                                    NaN
documents_task                                                   NaN
priority                                         Behavioural Failure
cause                                  JPC - Safety - Documentation
report_status                                                   Open
task_group                                                    Safety

                                                                   1  \
ref_form                                                 F145185.4
status_form                                                 Opened
location_form        01 Daily Site Diary>Site Management>JPC Projec...
name                                   1328 CM-SM-FRM-001 Site Diary
created_form                                              15/09/2020
type_form                                            Site Management
status_changed_form                                       15/09/2020
open_actions                                                       0
total_actions                                                      0
association_form                                                 NaN
overdue_form                                                   False
images_form                                                     True
comments_form                                                  False
documents_form                                                 False
project                                                         1328
report_forms_status                                             Open
report_forms_group                                   Site Management
ref_task                                                 T116412.200
```

```
status_task                                                   Closed
location_task        QC & BC(A)R>ITP 02 Architectural & M&E Service...
description                                                   Metsec
created_task                                              14/09/2020
target                                                           NaN
type_task                                        JPC - Progress Photo
to_package                                      Ceilings & Partitions
status_changed_task                                       14/09/2020
association_task                                                 NaN
overdue_task                                                   False
images_task                                                     True
comments_task                                                  False
documents_task                                                 False
priority                                                         NaN
cause                                                            NaN
report_status                                                 Closed
task_group                                           Site Management

                                                                   2  \
ref_form                                                 F145185.4
status_form                                                 Opened
location_form        01 Daily Site Diary>Site Management>JPC Projec...
name                                    1328 CM-SM-FRM-001 Site Diary
created_form                                            15/09/2020
type_form                                          Site Management
status_changed_form                                     15/09/2020
open_actions                                                     0
total_actions                                                    0
association_form                                               NaN
overdue_form                                                 False
images_form                                                   True
comments_form                                                False
documents_form                                               False
project                                                       1328
report_forms_status                                           Open
report_forms_group                                 Site Management
ref_task                                               T141663.27
status_task                                    EHS Good Observation
location_task        JPC Project Management>EHS Management>01 Inspe...
description          Good clear exclusion zones and access through ...
created_task                                            14/09/2020
target                                                         NaN
type_task                     Safety Notice (Green) - Good Observation
to_package                                         Main Contractor
status_changed_task                                     14/09/2020
association_task                                        FormAnswer
overdue_task                                                 False
images_task                                                   True
comments_task                                                False
documents_task                                               False
priority                                                       NaN
cause                                        JPC - Safety - Access
report_status                                               Closed
task_group                                                  Safety

                                                                   3  \
ref_form                                                     F145185.4
status_form                                                     Opened
location_form        01 Daily Site Diary>Site Management>JPC Projec...
name                                     1328 CM-SM-FRM-001 Site Diary
created_form                                                15/09/2020
type_form                                              Site Management
status_changed_form                                         15/09/2020
open_actions                                                         0
total_actions                                                        0
association_form                                                   NaN
overdue_form                                                     False
images_form                                                       True
comments_form                                                    False
documents_form                                                   False
project                                                           1328
report_forms_status                                               Open
report_forms_group                                     Site Management
ref_task                                                   T116412.199
status_task                                                     Closed
location_task        QC & BC(A)R>ITP 02 Architectural & M&E Service...
description                                                   RC walls
created_task                                                14/09/2020
target                                                             NaN
type_task                                         JPC - Progress Photo
to_package                                            Precast Concrete
status_changed_task                                         14/09/2020
association_task                                                   NaN
overdue_task                                                     False
images_task                                                       True
comments_task                                                    False
documents_task                                                   False
priority                                                           NaN
cause                                                              NaN
report_status                                                   Closed
task_group                                             Site Management

                                                                     4
ref_form                                                     F145185.4
status_form                                                     Opened
location_form        01 Daily Site Diary>Site Management>JPC Projec...
name                                     1328 CM-SM-FRM-001 Site Diary
created_form                                                15/09/2020
type_form                                              Site Management
status_changed_form                                         15/09/2020
open_actions                                                         0
total_actions                                                        0
association_form                                                   NaN
overdue_form                                                     False
images_form                                                       True
comments_form                                                    False
documents_form                                                   False
project                                                           1328
report_forms_status                                               Open
report_forms_group                                     Site Management
ref_task                                                    T141663.26
status_task                                       EHS Good Observation
```

```
location_task        JPC Project Management>EHS Management>01 Inspe...
description          block 02 working level has good housekeeping, ...
created_task                                                14/09/2020
target                                                             NaN
type_task                     Safety Notice (Green) - Good Observation
to_package                                            Precast Concrete
status_changed_task                                         14/09/2020
association_task                                            FormAnswer
overdue_task                                                     False
images_task                                                       True
comments_task                                                    False
documents_task                                                   False
priority                                                           NaN
cause                                     JPC - Safety - House Keeping
report_status                                                   Closed
task_group                                                      Safety
```

In [25]:

```python
# Compare similar columns between form and task
similar_cols = [
    ("status_form", "status_task"),
    ("location_form", "location_task"),
    ("type_form", "type_task"),
    ("created_form", "created_task")
]

for col1, col2 in similar_cols:
    same_pct = (merged_df[col1] == merged_df[col2]).mean() * 100
    print(f"{col1} vs {col2}: {same_pct:.2f}% identical")
```

```
status_form vs status_task: 15.29% identical
location_form vs location_task: 0.79% identical
type_form vs type_task: 0.00% identical
created_form vs created_task: 0.48% identical
```

In [27]:

```python
# Remove exact duplicate rows (if any)
dups = merged_df.duplicated().sum()
print("Exact duplicate rows:", dups)
if dups > 0:
    merged_df = merged_df.drop_duplicates()
    print("After drop_duplicates shape:", merged_df.shape)
```

```
Exact duplicate rows: 0
```

In [28]:

```python
# Identify potential join-duplicates (many-to-many) and check unique keys
# find columns that came from forms vs tasks (common suffixes _form/_task might exi
print([c for c in merged_df.columns if c.endswith('_form')][:20])
print([c for c in merged_df.columns if c.endswith('_task')][:20])

# distinct projects count
print("Unique projects:", merged_df['project'].nunique())
# how many rows per project (quick distribution)
print(merged_df['project'].value_counts().head(20))
```

```
['ref_form', 'status_form', 'location_form', 'created_form', 'type_form', 'status_ch
anged_form', 'association_form', 'overdue_form', 'images_form', 'comments_form', 'do
cuments_form']
['ref_task', 'status_task', 'location_task', 'created_task', 'type_task', 'status_ch
anged_task', 'association_task', 'overdue_task', 'images_task', 'comments_task', 'do
cuments_task']
Unique projects: 8
project
1328    15165293
1330     7916916
1335     1018668
1340      740280
1338      667080
1329      579336
1345      221760
1343      150876
Name: count, dtype: int64
```

In [29]:

```python
# Parse date columns robustly
# find likely date columns
date_cols = [c for c in merged_df.columns if ('created' in c or 'status_changed' in
print("Detected date cols:", date_cols)

# convert to datetime (dayfirst since your sample used dd/mm/yyyy)
for c in date_cols:
    merged_df[c] = pd.to_datetime(merged_df[c], dayfirst=True, errors='coerce')

# check how many parsed vs NaT
for c in date_cols:
    print(c, "=> nulls:", merged_df[c].isnull().sum())
```

```
Detected date cols: ['created_form', 'status_changed_form', 'created_task', 'status_
changed_task']
created_form => nulls: 0
status_changed_form => nulls: 0
created_task => nulls: 0
status_changed_task => nulls: 0
```

In [30]:

```python
# Create basic task-level features (close time, is_closed, is_overdue)
# pick sensible column names; adjust if your names differ
created_col = 'created_task' if 'created_task' in merged_df.columns else 'created'
status_changed_col = 'status_changed_task' if 'status_changed_task' in merged_df.co

# compute close duration if both exist
if created_col in merged_df.columns and status_changed_col in merged_df.columns:
    merged_df['close_duration_days'] = (merged_df[status_changed_col] - merged_df[c

# closed flag
if 'status_task' in merged_df.columns:
    merged_df['is_closed'] = merged_df['status_task'].fillna('').str.contains('clos

# overdue flag — use existing column if present, else infer from close duration + p
if 'overdue' not in merged_df.columns and 'close_duration_days' in merged_df.column
    # simple proxy: negative close_duration => suspicious; leave as NaN otherwise
    merged_df['overdue_inferred'] = merged_df['close_duration_days'] > 0
```

In [33]:

```python
# Compute KPIs separately on original tables (recommended), then join
# Standardize original tables too
forms = forms_df.copy()
tasks = tasks_df.copy()
forms.columns = forms.columns.str.strip().str.lower().str.replace(' ', '_')
tasks.columns = tasks.columns.str.strip().str.lower().str.replace(' ', '_')
tasks['project'] = tasks['project'].astype(str).str.strip().str.lower()
forms['project'] = forms['project'].astype(str).str.strip().str.lower()

# Parse dates in tasks
tasks['created'] = pd.to_datetime(tasks['created'], dayfirst=True, errors='coerce')
tasks['status_changed'] = pd.to_datetime(tasks['status_changed'], dayfirst=True, er
tasks['close_duration_days'] = (tasks['status_changed'] - tasks['created']).dt.days

# KPI A: completion rate per project (% closed)
completion_rate = (
    tasks.groupby('project')['status']
    .apply(lambda s: (s.str.lower() == 'closed').mean() * 100)
    .reset_index(name='completion_rate_pct')
)

# KPI B: overdue rate per project (% overdue)
if 'overdue' in tasks.columns:
    overdue_rate = tasks.groupby('project')['overdue'].mean().reset_index(name='ove
    overdue_rate['overdue_rate_pct'] *= 100
else:
    overdue_rate = tasks.groupby('project')['close_duration_days'].apply(lambda x:

# KPI C: avg close time per project (days)
avg_close = tasks.groupby('project')['close_duration_days'].mean().reset_index(name

# KPI D: open actions ratio from forms (open_actions / total_actions)
forms[['open_actions','total_actions']] = forms[['open_actions','total_actions']].a
open_ratio = forms.groupby('project').agg({'open_actions':'sum','total_actions':'su
open_ratio['open_actions_ratio'] = (open_ratio['open_actions'] / open_ratio['total_

# Merge KPI table
kpi = completion_rate.merge(overdue_rate, on='project', how='outer') \
                     .merge(avg_close, on='project', how='outer') \
                     .merge(open_ratio[['project','open_actions_ratio']], on='proje

kpi = kpi.fillna({'completion_rate_pct':0, 'overdue_rate_pct':0, 'avg_close_days':0
print(kpi)
kpi.to_csv("project_kpi_summary.csv", index=False)
```

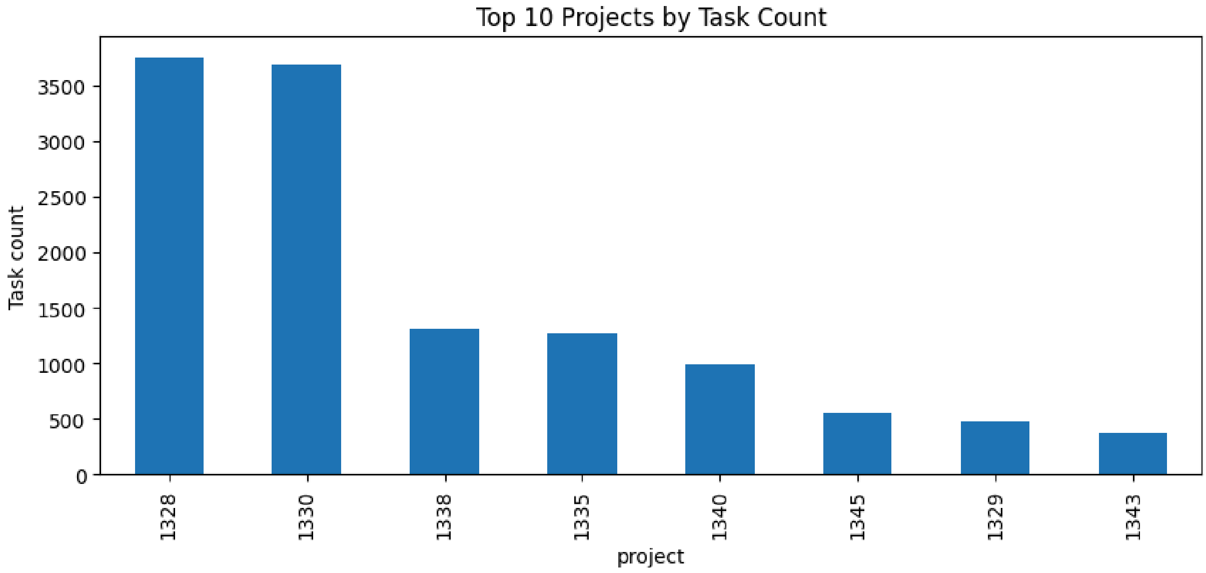
```
  project  completion_rate_pct  overdue_rate_pct  avg_close_days  \
0    1328            56.091709          5.465209       16.434018   
1    1329            51.464435          5.439331        5.853556   
2    1330            48.507058          6.134636       11.596906   
3    1335            43.804262         16.416732       10.158642   
4    1338            36.391437         14.678899        8.819572   
5    1340            66.834171          0.000000        2.152764   
6    1343            36.745407          0.000000        3.902887   
7    1345            28.928571          0.178571        3.848214   

   open_actions_ratio  
0            0.046180  
1            0.071090  
2            0.021220  
3            0.120635  
4            0.283255  
5            0.021322  
6            0.022409  
7            0.022409  
```

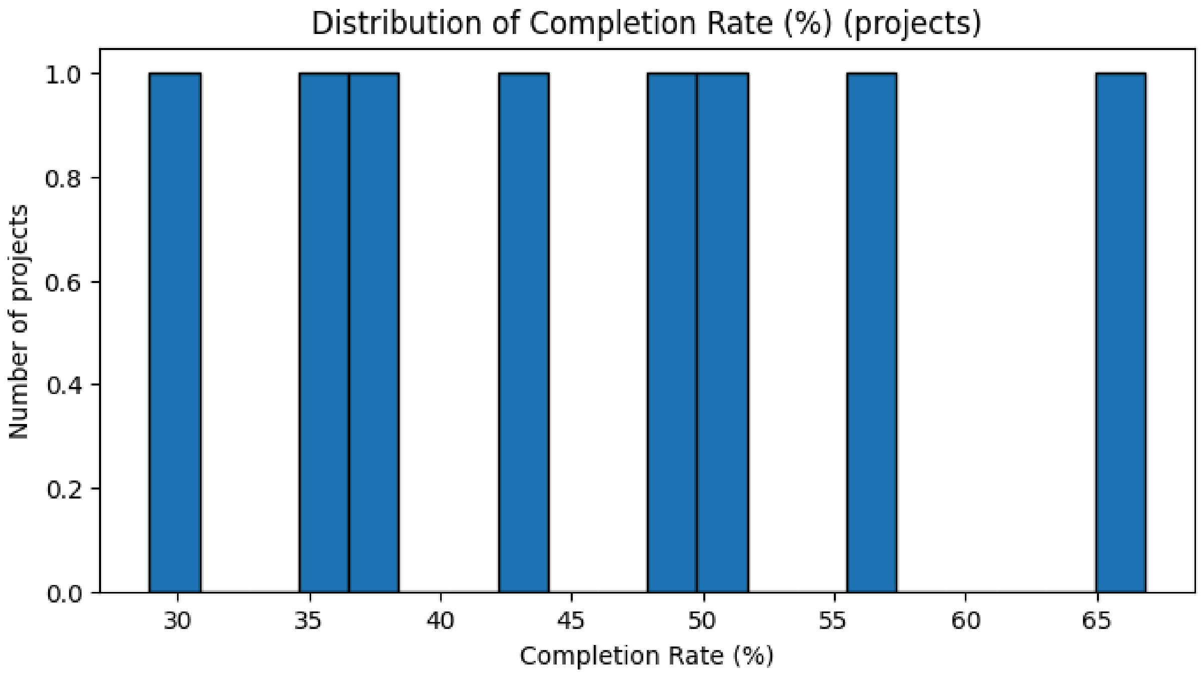
Quick EDA & visuals

In [34]:
```python
import matplotlib.pyplot as plt

# Top 10 projects by number of tasks
top_projects = tasks['project'].value_counts().nlargest(10)
top_projects.plot(kind='bar', figsize=(10,4), title='Top 10 Projects by Task Count'
plt.ylabel('Task count')
plt.show()
```

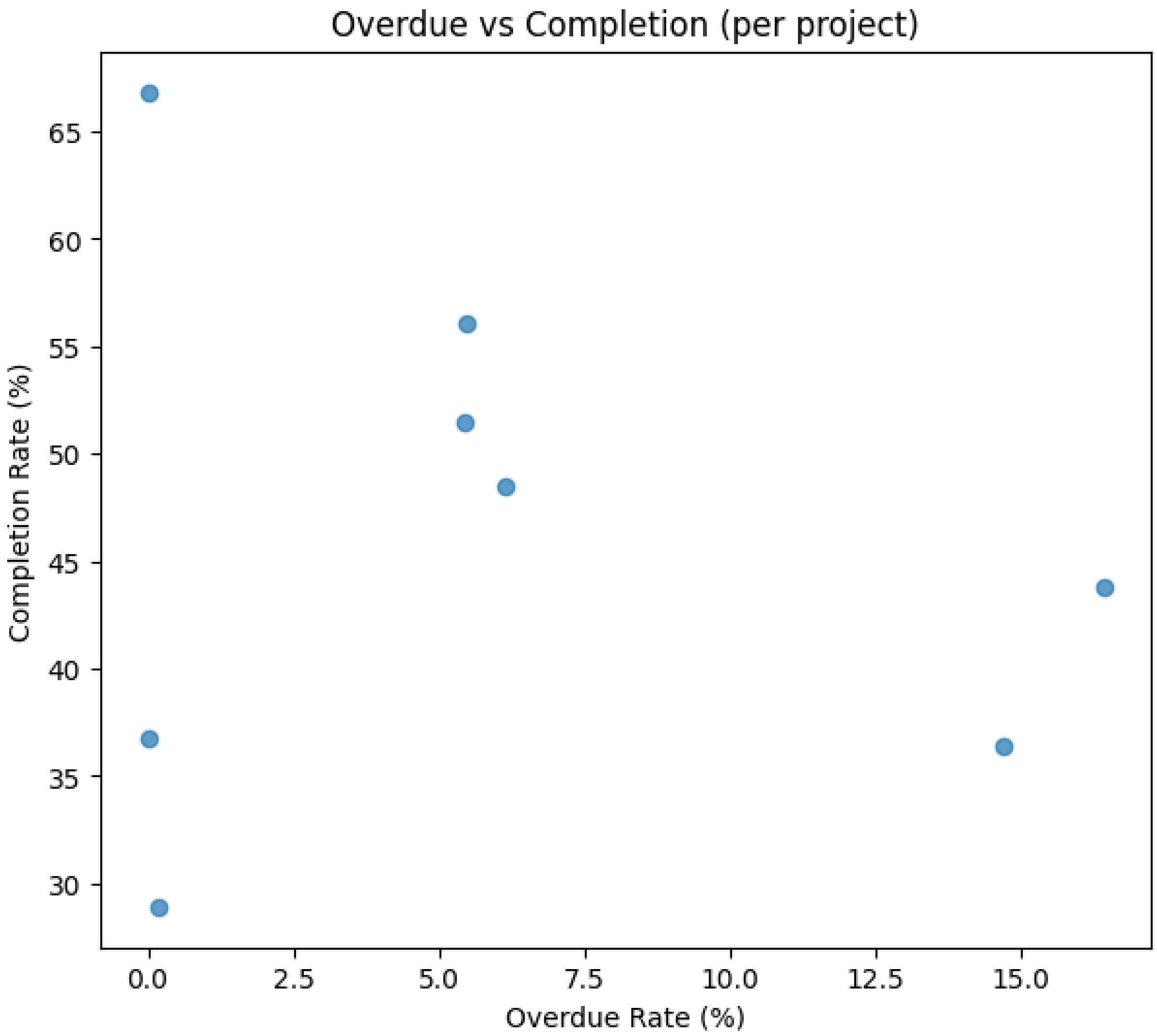


In [35]:
```python
# Completion rate distribution
plt.figure(figsize=(8,4))
plt.hist(kpi['completion_rate_pct'], bins=20, edgecolor='k')
plt.title('Distribution of Completion Rate (%) (projects)')
plt.xlabel('Completion Rate (%)'); plt.ylabel('Number of projects')
plt.show()
```



In [36]:

```
# Overdue vs completion scatter
plt.figure(figsize=(7,6))
plt.scatter(kpi['overdue_rate_pct'], kpi['completion_rate_pct'], alpha=0.7)
plt.xlabel('Overdue Rate (%)'); plt.ylabel('Completion Rate (%)')
plt.title('Overdue vs Completion (per project)')
plt.show()
```



Flag anomalies / priority list for intervention

In [38]:
```
# simple rules to flag projects requiring attention
kpi['flag_low_completion'] = kpi['completion_rate_pct'] < 60
kpi['flag_high_overdue'] = kpi['overdue_rate_pct'] > 30
kpi['needs_attention'] = kpi[['flag_low_completion','flag_high_overdue']].any(axis=
```

In [39]:
```
# list top flagged projects
attention = kpi[kpi['needs_attention']].sort_values(['overdue_rate_pct','completion
print("Projects needing attention:", attention.head(20))
```

```
Projects needing attention:   project  completion_rate_pct  overdue_rate_pct  avg_cl
ose_days  \
3    1335            43.804262        16.416732       10.158642
4    1338            36.391437        14.678899        8.819572
2    1330            48.507058         6.134636       11.596906
0    1328            56.091709         5.465209       16.434018
1    1329            51.464435         5.439331        5.853556
7    1345            28.928571         0.178571        3.848214
6    1343            36.745407         0.000000        3.902887

   open_actions_ratio  flag_low_completion  flag_high_overdue  needs_attention
3            0.120635                 True              False             True
4            0.283255                 True              False             True
2            0.021220                 True              False             True
0            0.046180                 True              False             True
1            0.071090                 True              False             True
7            0.022409                 True              False             True
6            0.022409                 True              False             True
```

Save outputs & prepare deliverables

In [ ]:

```python
# save cleaned merged and KPI files
merged_df.to_csv("merged_cleaned.csv", index=False)
kpi.to_csv("project_kpi_summary.csv", index=False)
```

In [ ]: