cricket-data-modification

November 24, 2019

[ ]:
```python
import pandas as pd
```

[ ]:
```python
df_info=pd.read_csv(r'./full/odi_info.csv')
```

[ ]:
```python
df_scorecard=pd.read_csv(r'./full/odi_scorecard.csv')
```

[ ]:
```python
df_scorecard = df_scorecard.astype({"match-id": str,'innings':str})
```

0.0.1 Delete Attributes

[ ]:
```python
df_info=df_info.drop(['eliminator','date-1','date-2','date-3','date-4'],axis=1)
```

[ ]:
```python
df_info=df_info.drop(['winner-innings'],axis=1)
```

[ ]:
```python
df_info=df_info.drop(['index'],axis=1)
```

0.0.2 Delete matches

[ ]:
```python
for i in df_info['match-id'].index:
    if '(' in df_info['match-id'][i]:
        print(i)
```

[ ]:
```python
df_info.drop(df_info[df_info['match-id']=='300438 (1)'].index,inplace=True)
df_info.drop(df_info[df_info['match-id']=='812777 (1)'].index,inplace=True)
```

[ ]:
```python
df_info.drop(df_info[df_info['match-id']=='1050229'].index,inplace=True)
```

[ ]:
```python
df_info.drop(df_info[df_info['match-id']=='1022599'].index,inplace=True)
```

[ ]:
```python
df_info.drop(df_info[df_info['match-id']=='426423'].index,inplace=True)
```

[ ]:
```python
df_info.drop(df_info[df_info['match-id']=='915773'].index,inplace=True)
```

[ ]:
```python
df_scorecard.drop(df_scorecard[df_scorecard['match-id']=='1050229'].
 ↪index,inplace=True)
```

[ ]:
```python
df_scorecard.drop(df_scorecard[df_scorecard['match-id']=='1022599'].
 ↪index,inplace=True)
```

[ ]:
```python
df_scorecard.drop(df_scorecard[df_scorecard['match-id']=='426423'].
 ↪index,inplace=True)
```

[ ]:

```
df_scorecard.drop(df_scorecard[df_scorecard['match-id']=='915773'].
 ↪index,inplace=True)
```

[ ]:

```
df_scorecard.drop(df_scorecard[df_scorecard['match-id']=='300438 (1)'].
 ↪index,inplace=True)
df_scorecard.drop(df_scorecard[df_scorecard['match-id']=='812777 (1)'].
 ↪index,inplace=True)
```

[ ]:

```
match_id=df_info[(df_info['competition']=='Indian Premier League') |␣
 ↪(df_info['competition']=='ICC World Twenty20') |␣
 ↪(df_info['competition']=='Big Bash League')]['match-id']
for i in match_id:
    df_scorecard.drop(df_scorecard[df_scorecard['match-id']==i].
 ↪index,inplace=True)
    df_info.drop(df_info[df_info['match-id']==i].index,inplace=True)
```

[ ]:

```
for i in match_id:
    df_scorecard.drop(df_scorecard[df_scorecard['match-id']==i].
 ↪index,inplace=True)
    df_info.drop(df_info[df_info['match-id']==i].index,inplace=True)
```

[ ]:

```
# df_scorecard.drop(['Unnamed: 0','Unnamed: 0.1'],axis=1,inplace=True)
```

0.0.3 Additional Formatting

[ ]:

```
df_zero=df_scorecard[df_scorecard['innings']=='0']
for i, v in df_zero.iterrows():
    print(v['match-id'])
```

[ ]:

```
df_info['competition']=df_info['competition'].apply(lambda x: str(x).
 ↪replace(r'"',''))
```

[ ]:

```
df_info['venue']=df_info['venue'].apply(lambda x: str(x).replace(r'"',''))
```

[ ]:

```
df_info['year']=df_info['date'].apply(lambda x: str(x).split(r'/')[0])
```

0.0.4 Aggregate scorecard

[ ]:

```
df_scorecard_agg=df_scorecard.groupby(['match-id','team-name'],as_index=False).
 ↪sum()
```

[ ]:

```
match_id=df_scorecard_agg[df_scorecard_agg['runs-scored']==0]['match-id']
```

[ ]:

```
cancelled_matches=df_scorecard_agg[(df_scorecard_agg['runs-scored']<=50) &␣
 ↪(df_scorecard_agg['batting-position']<66) &␣
 ↪(df_scorecard_agg['balls-played']<300)]
match_id=match_id.
 ↪append(cancelled_matches[cancelled_matches['runs-scored']<cancelled_matches['runs-given']][
```

[ ]:

```
for i in match_id:
    df_scorecard.drop(df_scorecard[df_scorecard['match-id']==i].
 ↪index,inplace=True)
    df_info.drop(df_info[df_info['match-id']==i].index,inplace=True)
    df_scorecard_agg.drop(df_scorecard_agg[df_scorecard_agg['match-id']==i].
 ↪index,inplace=True)
```

0.0.5 Save files

[ ]:

```
df_info.sort_values(['match-id'],inplace=True)
df_info.to_csv(r'./full/odi_info.csv',index=False)
```

[ ]:

```
df_scorecard.sort_values(['match-id'],inplace=True)
df_scorecard.to_csv(r'./full/odi_scorecard.csv',index=False)
```

[ ]:

```
df_scorecard_agg.sort_values(['match-id'],inplace=True)
df_scorecard_agg.to_csv(r'./full/odi_scorecard_agg.csv',index=False)
```