In [1]:

```
import pandas as pd
from code.preprocessing import *
from code.analytics import *
from code.visualization import *
from code.modeling import *
```

In [2]:

```
match_results_df = pd.read_json('../data/raw/match_results.json')
innings_results_df = pd.read_json('../data/raw/innings_results.json')
innings_results_df = cleanse_innings_results(innings_results_df, match_results_df)
print(innings_results_df)
innings_results_df.to_csv('../data/clean/innings_results.csv')
```

```
           matchid       team  innings  over  runs.total wicket.kind
0        1000887.0  Australia        1   0.1           0         NaN
1        1000887.0  Australia        1   0.2           0         NaN
2        1000887.0  Australia        1   0.3           0         NaN
3        1000887.0  Australia        1   0.4           0         NaN
4        1000887.0  Australia        1   0.5           1         NaN
...            ...        ...      ...   ...         ...         ...
1150196   997995.0   Scotland        2  46.6           0         NaN
1150197   997995.0   Scotland        2  47.1           0         NaN
1150198   997995.0   Scotland        2  47.2           0         NaN
1150199   997995.0   Scotland        2  47.3           0         NaN
1150200   997995.0   Scotland        2  47.4           4         NaN

[1124215 rows x 6 columns]
```

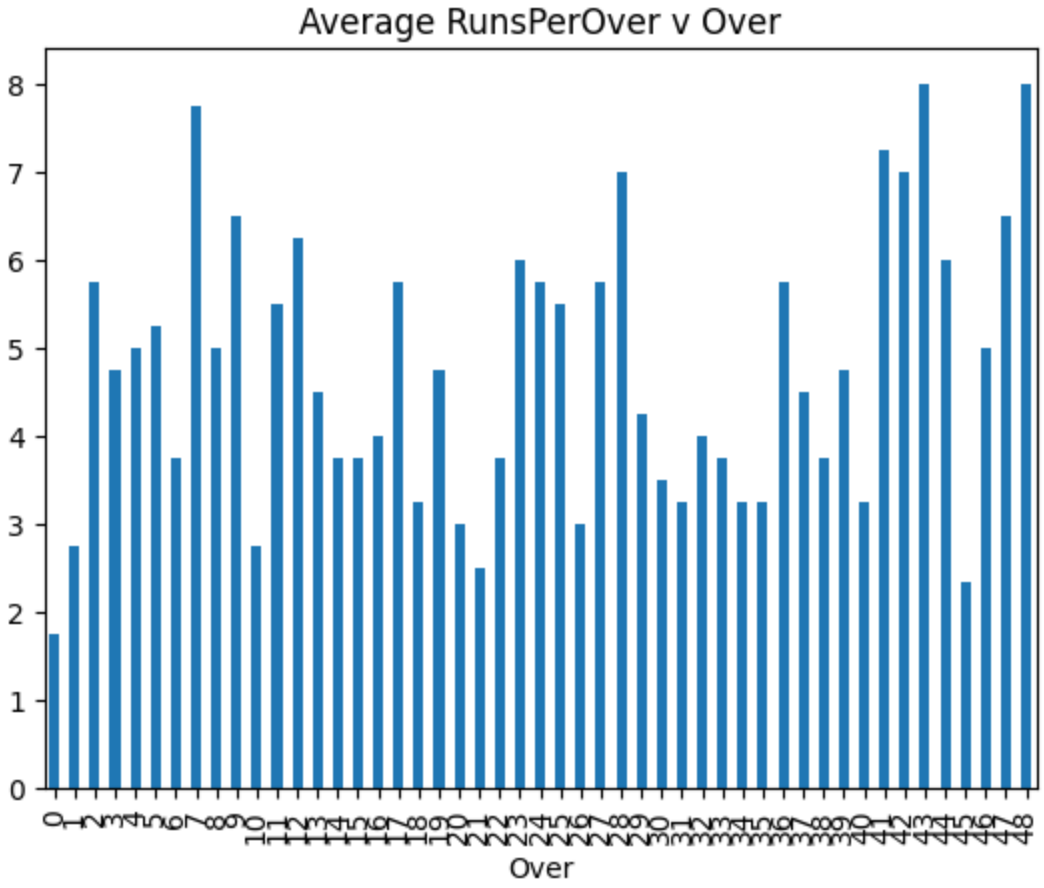
In [3]:

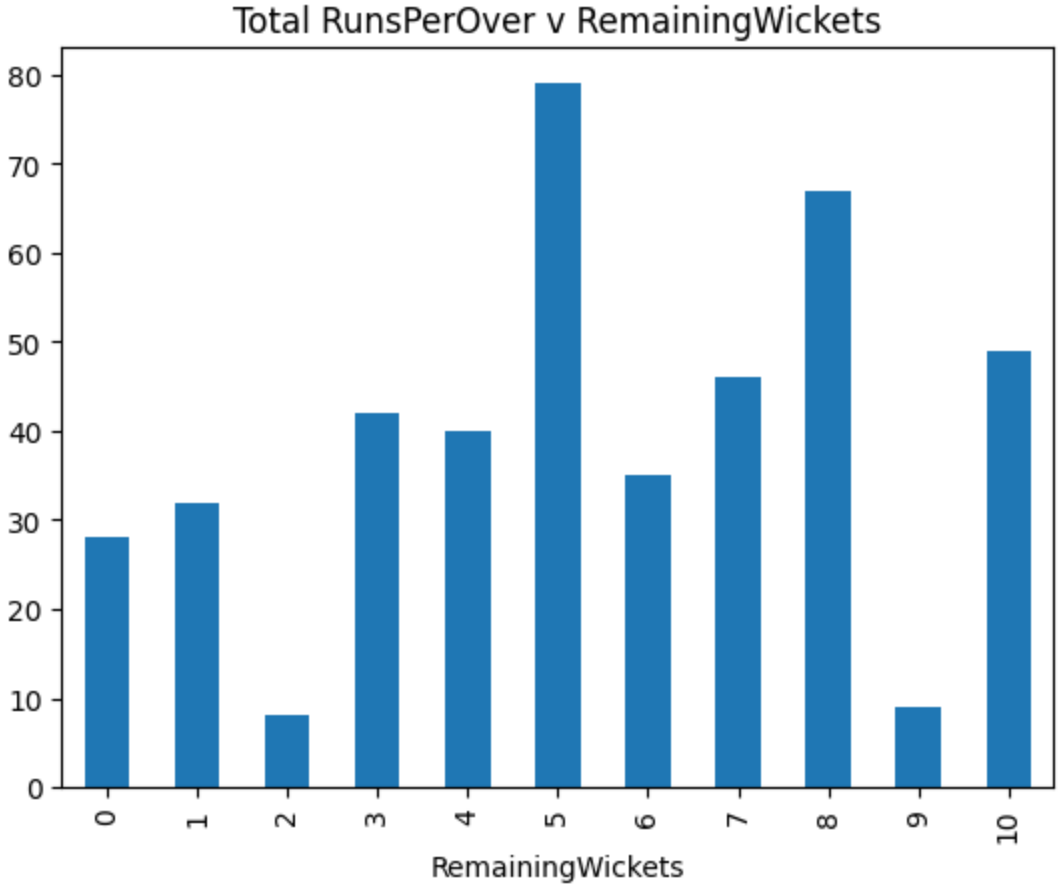
```
innings_results_df = pd.read_csv('../data/clean/innings_results.csv', index_col=0, low_mer
overs_summary_df = assemble_over_statistics(innings_results_df)
print(overs_summary_df)
overs_summary_df.to_csv('../data/results/over_summary.csv')
```

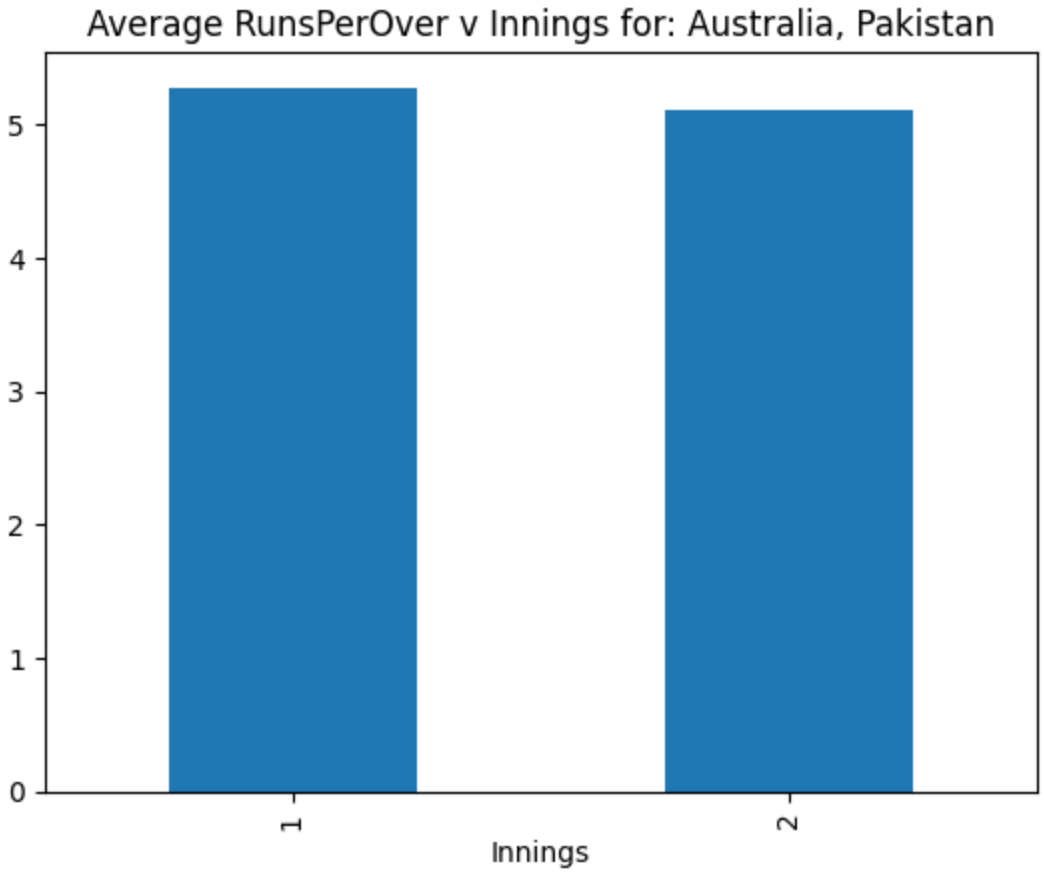
```
        index    matchid       Team Innings Over RemainingOvers RunsPerOver  \
0           0  1000887.0  Australia       1    0             50           1
1           0  1000887.0  Australia       1    1             49           1
2           0  1000887.0  Australia       1    2             48           3
3           0  1000887.0  Australia       1    3             47           6
4           0  1000887.0  Australia       1    4             46           2
...       ...        ...        ...     ...  ...            ...         ...
179169      0   997995.0   Scotland       2   43              7           4
179170      0   997995.0   Scotland       2   44              6           2
179171      0   997995.0   Scotland       2   45              5           9
179172      0   997995.0   Scotland       2   46              4           7
179173      0   997995.0   Scotland       2   47              3           4

       TotalRuns RemainingWickets
0              1               10
1              2               10
2              5               10
3             11               10
4             13                8
...          ...              ...
179169       207                8
179170       209                7
179171       218                7
179172       225                7
179173       229                7

[179174 rows x 9 columns]
```

In [4]:

```
overs_summary_df = pd.read_csv('../data/results/over_summary.csv', index_col=0)
plot_runs_per_over(overs_summary_df, [1000887, 1000889], aggr="mean")
plot_runs_per_over(overs_summary_df, 1000887, target_col="RemainingWickets")
plot_runs_per_over(overs_summary_df, team=["Australia", "Pakistan"], aggr="mean", target_c
```


Average RunsPerOver v Over
Over


Total RunsPerOver v RemainingWickets
RemainingWickets



In [5]:
```
predict_runs_per_over(overs_summary_df)
```

```
Coefficients:
 [ 0.07230627 -1.39089275  1.0148446 ]
Mean squared error: 11.52
Coefficient of determination: 0.073
```

In [ ]: