Game Durations Across Regions

April 24, 2022

The duration of a match helps us infer the play style of the players of a region. Shorter match times may point at a more aggressive compared to a more calculated and slower play style. Different regions of the world differ in this metric in a lot of online games.

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
from scipy import stats
```

[2]:
```
matches = pd.read_csv('datasets/dota-2-matches/match.csv')

matches.head()
```

[2]:

	match_id	start_time	duration	tower_status_radiant	tower_status_dire	barracks_status_dire	barracks_status_radiant	first_blood_time	game_mode	radiant_win	negative_votes	positive_votes	cluster
0	0	1446750112	2375	1982	4	3	63	1	22	True	0	1	155
1	1	1446753078	2582	0	1846	63	0	221	22	False	0	2	154
2	2	1446764586	2716	256	1972	63	48	190	22	False	0	0	132
3	3	1446765723	3085	4	1924	51	3	40	22	False	0	0	191
4	4	1446796385	1887	2047	0	0	63	58	22	True	0	0	156

[3]:
```
# Convert duration from seconds to minutes for readability
matches['duration'] = matches['duration'] / 60
```

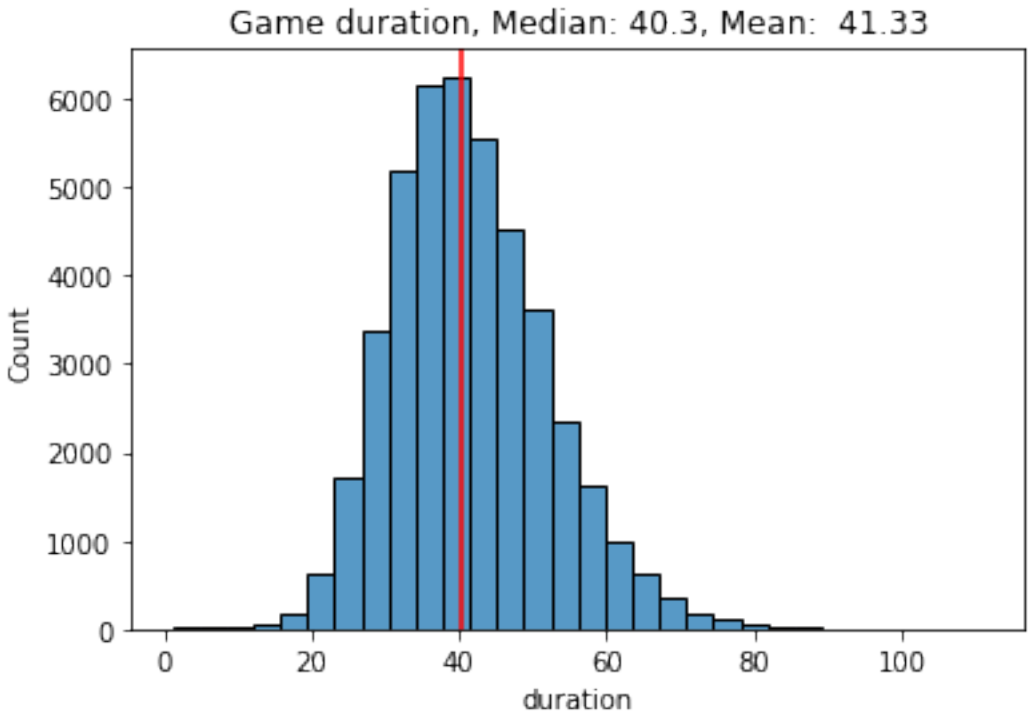
```
# Convert Unix timestamp to datetime
matches['date'] = pd.to_datetime(matches['start_time'], unit = 's')

# Merge the clusters table into the matches
clusters = pd.read_csv('datasets/dota-2-matches/cluster_regions.csv')
matches =  matches.merge(clusters, on = 'cluster', how = 'left')
matches = matches[['duration', 'game_mode', 'date', 'region']]

matches.head()
```

[3]:

	duration	game_mode	date	region
0	39.583333	22	2015-11-05 19:01:52	SINGAPORE
1	43.033333	22	2015-11-05 19:51:18	SINGAPORE
2	45.266667	22	2015-11-05 23:03:06	EUROPE
3	51.416667	22	2015-11-05 23:22:03	AUSTRIA
4	31.450000	22	2015-11-06 07:53:05	SINGAPORE

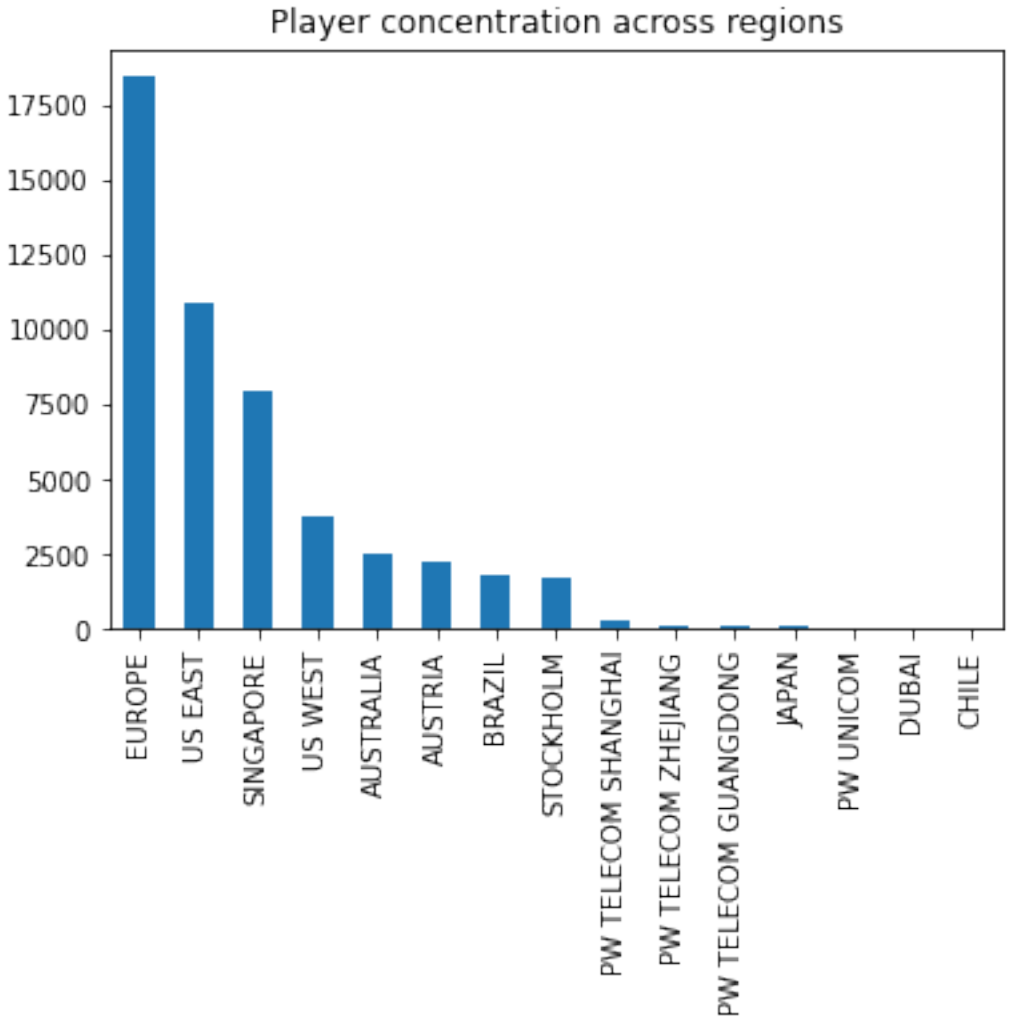
[15]:
```
sns.histplot(data=matches, x='duration', bins=30)
median_duration = matches['duration'].median()
mean_duration = matches['duration'].mean()

plt.axvline(x=median_duration, color='red')
plt.title(f"Game duration, Median: {median_duration}, Mean: {mean_duration: 0.
↪2f}")
```

[15]: Text(0.5, 1.0, 'Game duration, Median: 40.3, Mean: 41.33')



[5]:
```
matches['region'].value_counts().plot(kind = 'bar').set_title('Player␣
↪concentration across regions')
```

[5]:
```
Text(0.5, 1.0, 'Player concentration across regions')
```

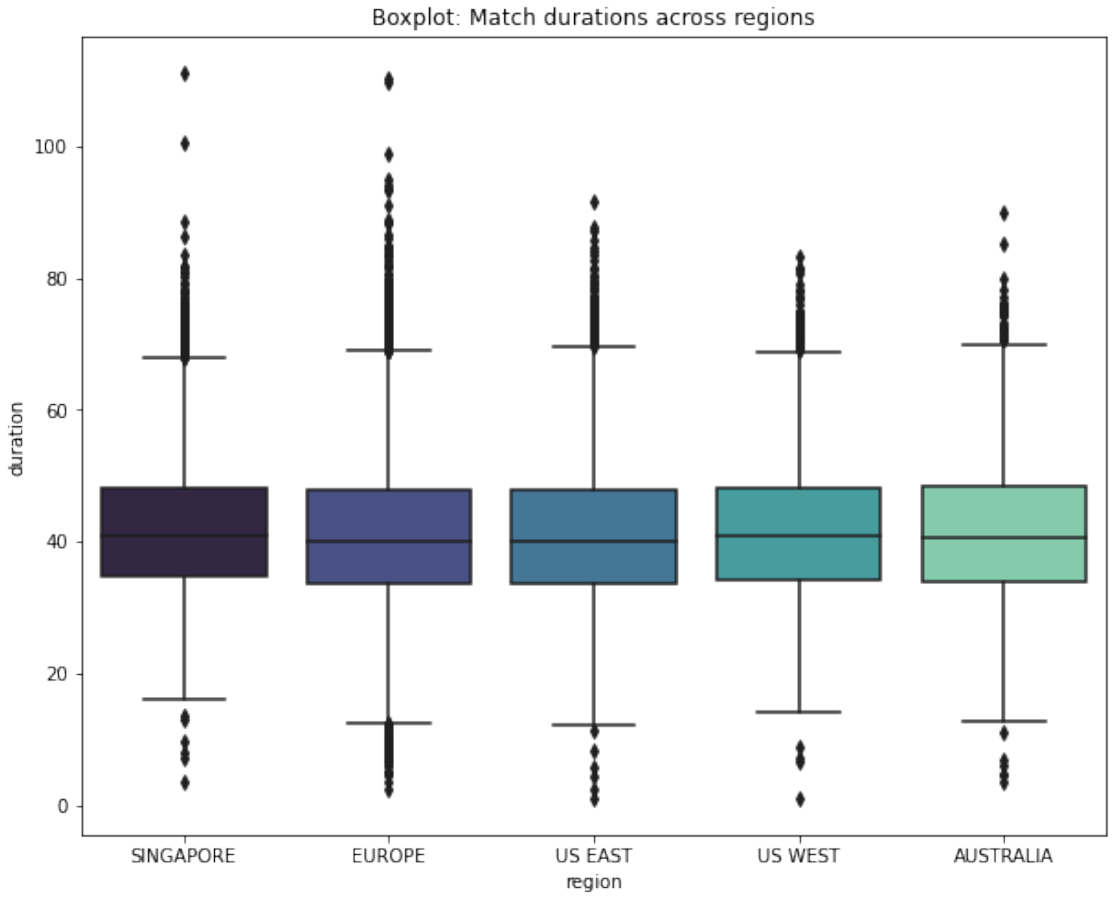


Selecting the top regions for analysis

[6]:
```
# Drop any region outside of the ones we are interested in
regions_to_keep = ['EUROPE', 'US EAST', 'SINGAPORE', 'US WEST', 'AUSTRALIA']

matches.drop(matches[ ~matches['region'].isin(regions_to_keep)].index, inplace␣
↪= True)
```

[7]:
```
plt.figure(figsize=(10,8))
plt.title("Boxplot: Match durations across regions")
sns.boxplot(x = 'region', y = 'duration', data=matches, palette='mako')
```

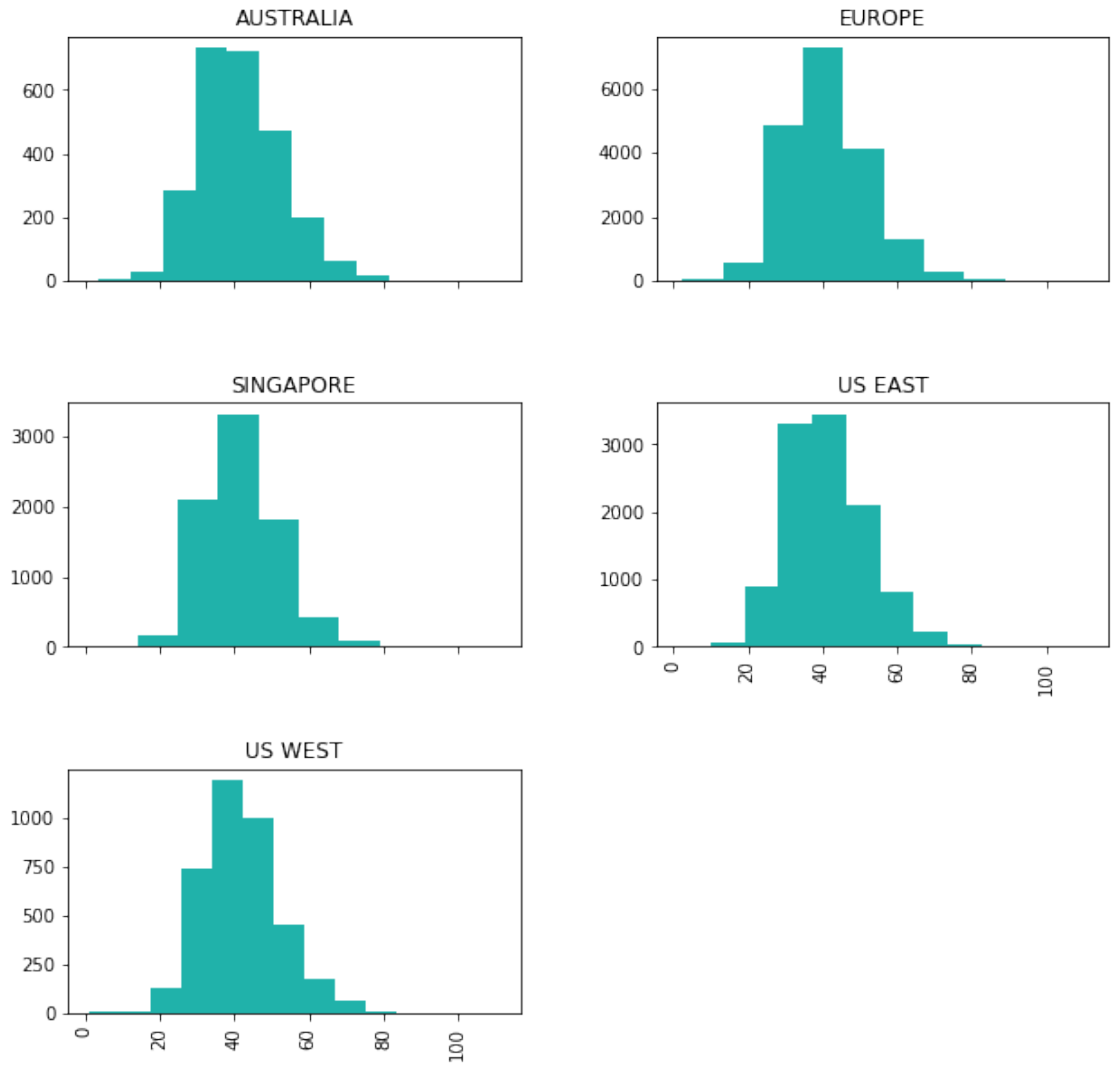
[7]:
```
<AxesSubplot:title={'center':'Boxplot: Match durations across regions'},
xlabel='region', ylabel='duration'>
```



Distribution of match durations across regions

[8]:
```
print('--- Match Duration Globally ---')
matches['duration'].hist(by = matches['region'], figsize = (10,10),␣
 ↪sharex=True, color='lightseagreen')
plt.show()
```

```
--- Match Duration Globally ---
```



[9]:
```
matches.groupby('region')['duration'].mean()
```

[9]:
```
region
AUSTRALIA    41.542880
EUROPE       41.083155
SINGAPORE    41.815132
US EAST      41.191995
US WEST      41.773940
Name: duration, dtype: float64
```

We can see minor differences in the distribution of durations across regions. To test the statistical significance of these results, we can perform a pairwise t-test for the following hypothesis:

H_o: The duration distribution means across regions are the same

H_a: The durations across regions are different

[10]:
```
import itertools

durations_by_region = []
for region in regions_to_keep:
    durations_by_region.append(matches[matches['region'] == region]['duration'])

regions = itertools.permutations(regions_to_keep, 2)

regions_duration = pd.DataFrame(columns=['region_1', 'region_2', 't_statistic',␣
 ↪'p_value'])

for r1, r2 in regions:
    max_len = min(len(matches[matches['region'] == r1]),␣
 ↪len(matches[matches['region'] == r2]))
    T, p = stats.ttest_ind(matches[matches['region'] == r1]['duration'][:
 ↪max_len],
                           matches[matches['region'] == r2]['duration'][:
 ↪max_len], equal_var=False)
    regions_duration.loc[len(regions_duration.index)] = [r1, r2, T, p]
```

Get regions with statistically significant mean values

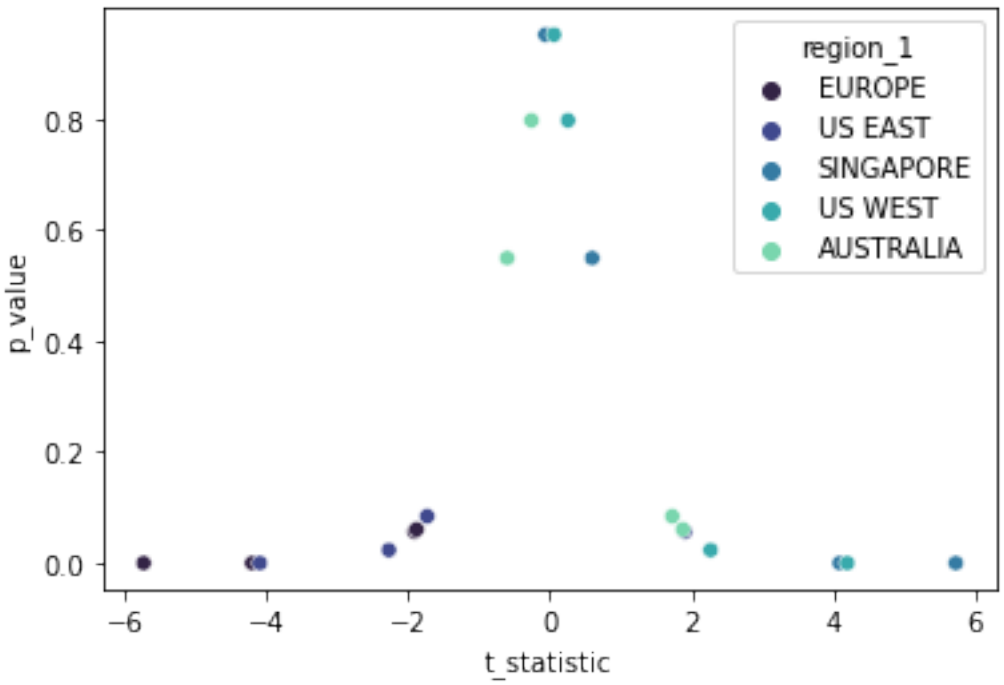
[27]:
```
regions_duration.iloc[[0,1,2,3,5,6,7,10,11,15],:]
```

[27]:

	region_1	region_2	t_statistic	p_value
0	EUROPE	US EAST	-1.906170	5.664125e-02
1	EUROPE	SINGAPORE	-5.722863	1.066483e-08
2	EUROPE	US WEST	-4.193992	2.772270e-05
3	EUROPE	AUSTRALIA	-1.875495	6.078261e-02
5	US EAST	SINGAPORE	-4.084156	4.445571e-05
6	US EAST	US WEST	-2.267031	2.341653e-02
7	US EAST	AUSTRALIA	-1.726790	8.426685e-02
10	SINGAPORE	US WEST	-0.061228	9.511791e-01
11	SINGAPORE	AUSTRALIA	0.599267	5.490225e-01
15	US WEST	AUSTRALIA	0.257338	7.969283e-01

[12]:
```
sns.scatterplot(x='t_statistic', y='p_value', data=regions_duration,␣
 ↪hue='region_1', palette='mako')
```

[12]:
```
<AxesSubplot:xlabel='t_statistic', ylabel='p_value'>
```



As we can see, the deviation is most significant for `SINGAPORE` and `AUSTRALIA` regions compared to other regions. We can infer that match durations are significantly longer for players in the `AUSTRALIA` and `SINGAPORE` region than any other top region in the world, which could mean they opt for a more slow and calculated strategy compared to the rest of the world. Compared to that, the duration for `EUROPE` and `US EAST` regions is the lowest, hinting towards a faster play style.