Part 1 - ValidateNormality

September 2, 2020

0.0.1 Intro

The objective of this notebook is to verify whether or not scoring throughout an NBA game could feasibly be simulated using Brownian Motion, in order to estimate win probability. This is accomplished by verifying whether minute-by-minute scoring changes in games follows a normal distribution. After I came up with the idea, I was also able to find a paper that I could reference on the general approach: https://www.stat.berkeley.edu/~aldous/157/Papers/stern.pdf

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

pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

[2]:
```
from nba_win_probability import dataloader, transformations, plotting
```

0.0.2 Load and Transform

Load and transform data in order to extract the minute-by-minute score differences in order to validate normality assumption.

[3]:
```
df = dataloader.load_season("2018")
transformed_df = transformations.transform_data_for_analysis(df)
transformed_df = transformed_df[['GAME_ID', 'PLAYER1_TEAM_ABBREVIATION',␣
↪'PLAYER2_TEAM_ABBREVIATION',  'PERIOD', 'SCORE', 'SCOREMARGIN', 'SEASON',␣
↪'QUARTER_TS', 'TIME_ELAPSED', 'MINUTE', 'SCORE_BY_MINUTE']]
transformed_df.head(5)
```

[3]:

	GAME_ID	PLAYER1_TEAM_ABBREVIATION	PLAYER2_TEAM_ABBREVIATION	PERIOD	SCORE	SCOREMARGIN	SEASON	QUARTER_TS	TIME_ELAPSED	MINUTE
0	21800001	BOS	PHI	1	NaN	0.0	2018	0.000000	0.000000	0.0
1	21800001	PHI	PHI	1						
2	21800001	BOS	BOS	1						
3	21800001	PHI	BOS	1						
4	21800001	BOS	BOS	1						

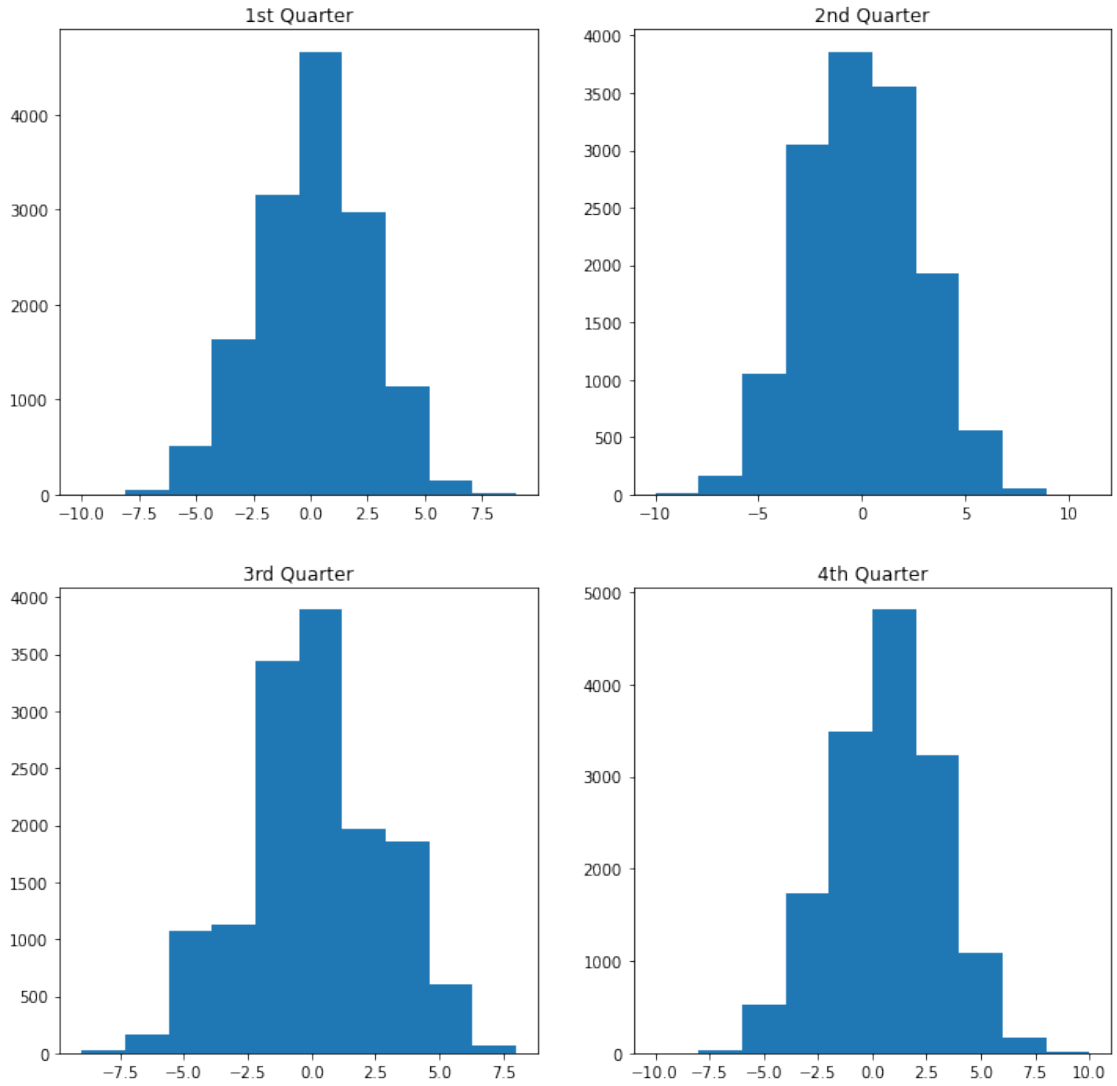
```
1  2 - 0         -2.0   2018    1.083333      1.083333     1.0
2  2 - 2          0.0   2018    2.683333      2.683333     2.0
3  6 - 4         -2.0   2018    3.750000      3.750000     3.0
4  8 - 7         -1.0   2018    4.566667      4.566667     4.0

   SCORE_BY_MINUTE
0              0.0
1             -2.0
2              2.0
3             -2.0
4              1.0
```

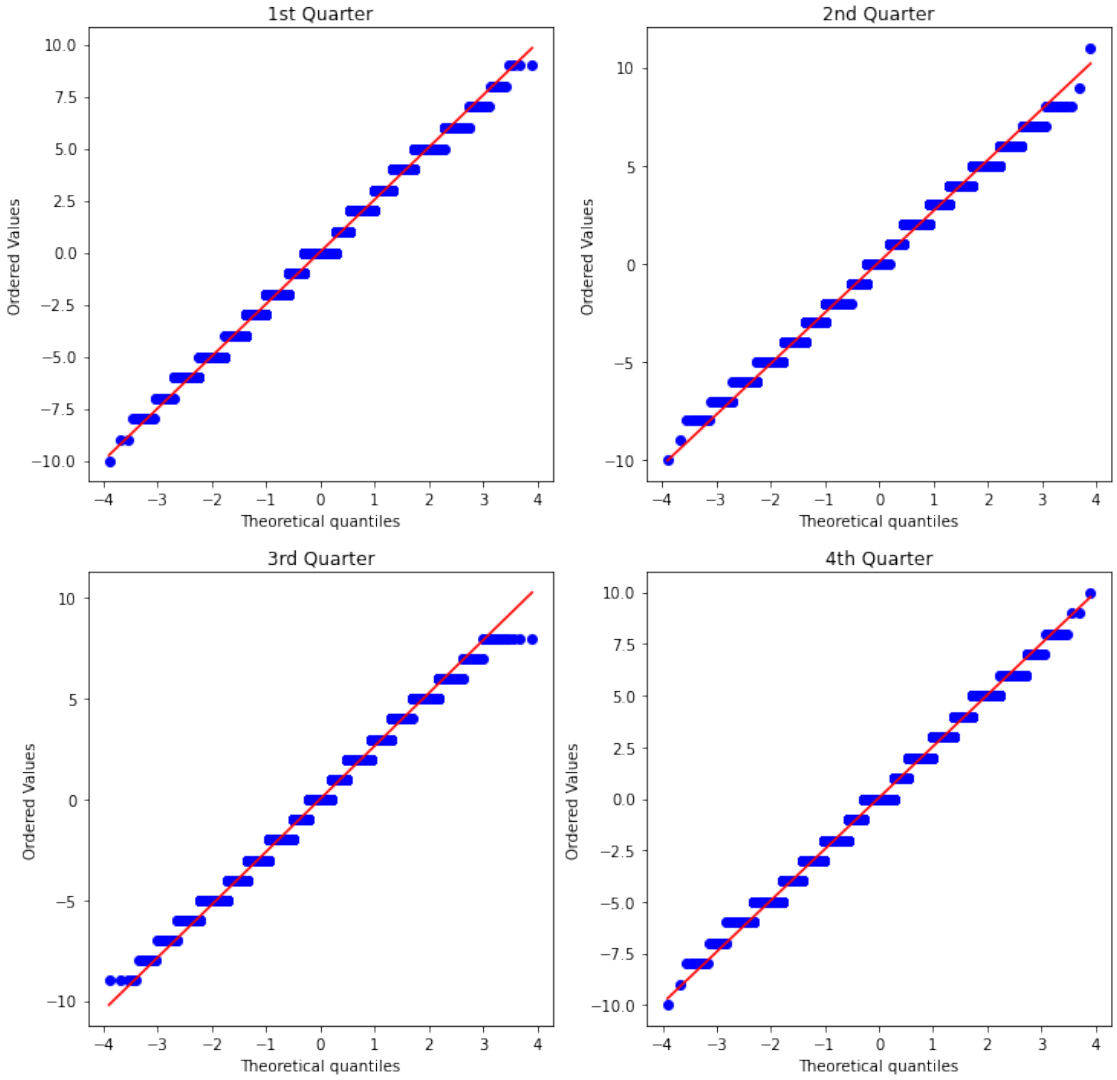
0.0.3 Normality Plots

The QQ plots and histograms below illustrate the normality of the data, on a quarter-by-quarter basis.

[4]:
```
plotting.histogram_plot_score_difference_by_quarter(transformed_df)
```



[5]:
```
plotting.qq_plot_score_difference_by_quarter(transformed_df)
```



0.0.4 Conclusion

The data here is fairly normal based on the histograms and the Q-Q plots. Obviously since the data is integer-based instead of continuous, there are a lot of repeated values across the board, which is noticeable in the Q-Q plot. The apparent normality makes NBA games a good candidate for simulation using Brownian Motion.

[ ]: