Movie Database Project

The Business Questions

- What is the most popular genre of movie?
- When is the best time to release a movie?

Data Scraping

1. Connected to The Movie Database's API
2. Downloaded data for films released in 2018 in two datasets - one qualitative, one quantitative
3. Merged both datasets into a csv
4. Focused analysis on two dependent variables: average voter rating and revenue

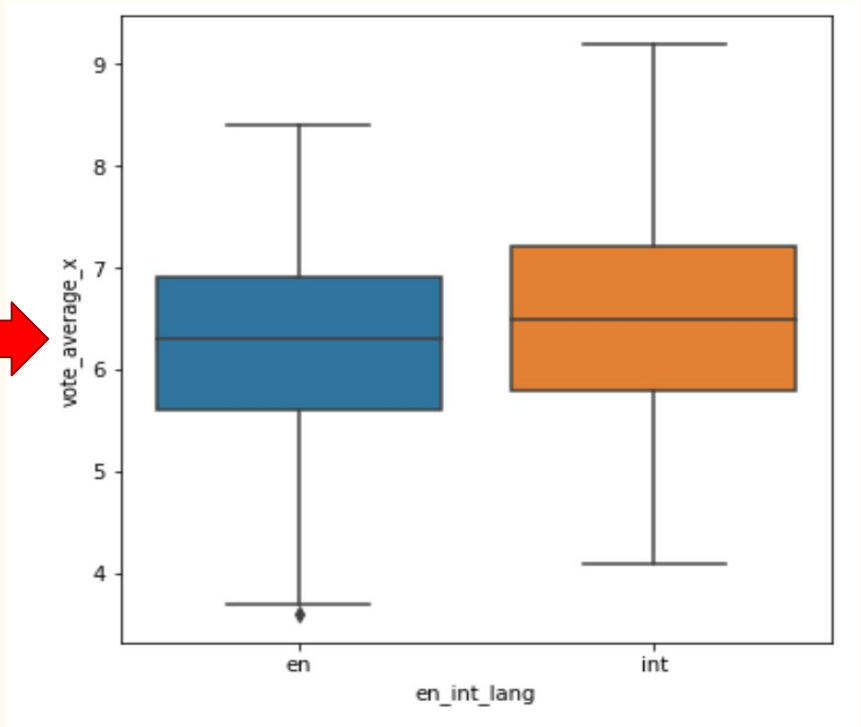
English Language vs International - EDA

English accounted for 58% of all movies

Grouped all other languages into ‘International’

Indication of a possible difference in Average Vote Score between English language movies and International language movies.

en	0.587
fr	0.124
it	0.062
es	0.050
de	0.026
hi	0.022
ja	0.019
pt	0.015
ru	0.013
ko	0.013
zh	0.012
ta	0.008
pl	0.005
nl	0.005
tr	0.005
id	0.003
da	0.003
sv	0.003
no	0.003
is	0.002
hu	0.002
fi	0.001
eu	0.001
sw	0.001
uk	0.001
tl	0.001
fa	0.001
cn	0.001
kn	0.001
ar	0.001
he	0.001



English Language vs International - Hypothesis

H0 Null Hypothesis:

There is no significant difference in user rating between English language films and non English language films.

Mean rating En = Mean rating Int

HA Alternative Hypothesis

There is a significant difference in user rating between English language films and non English language films

Mean rating En ≠ Mean rating Int

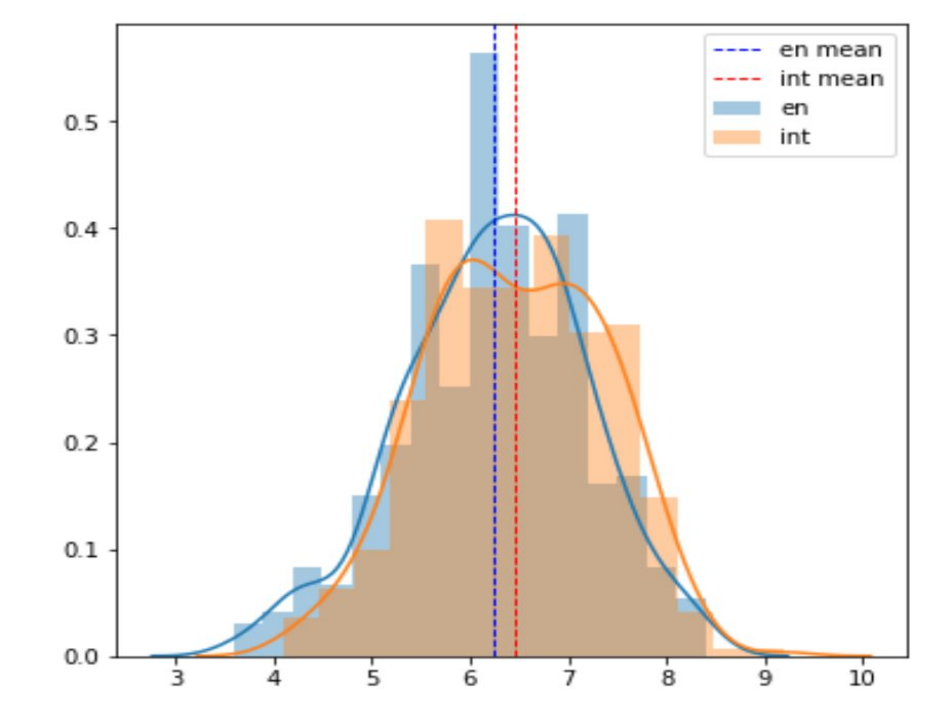
English Language vs International - Findings

Sample english = **556**, Sample interntn'l = **390**

Alpha = **0.05**
Welch's t-test= **-3.5230**
p-value = **0.00044948**
Cohen's d = **-0.23,** Power = **0.94**

Statistically significant difference BUT...
tiny effect size.

Recommendation: DO NOT invest substantial resources in developing a non-English movie as the actual pay off may not justify the resources spent.

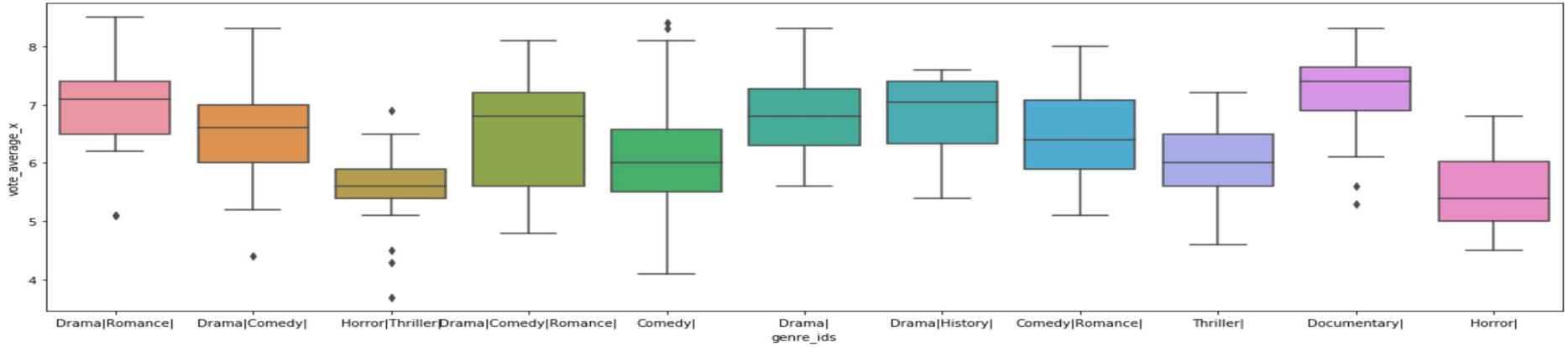


Genre - EDA

Many, many different movie genres released in 2018

Focussed on the top 10 genres based on number of movies released

Possible indication of difference in vote scores across the top 10 genres...

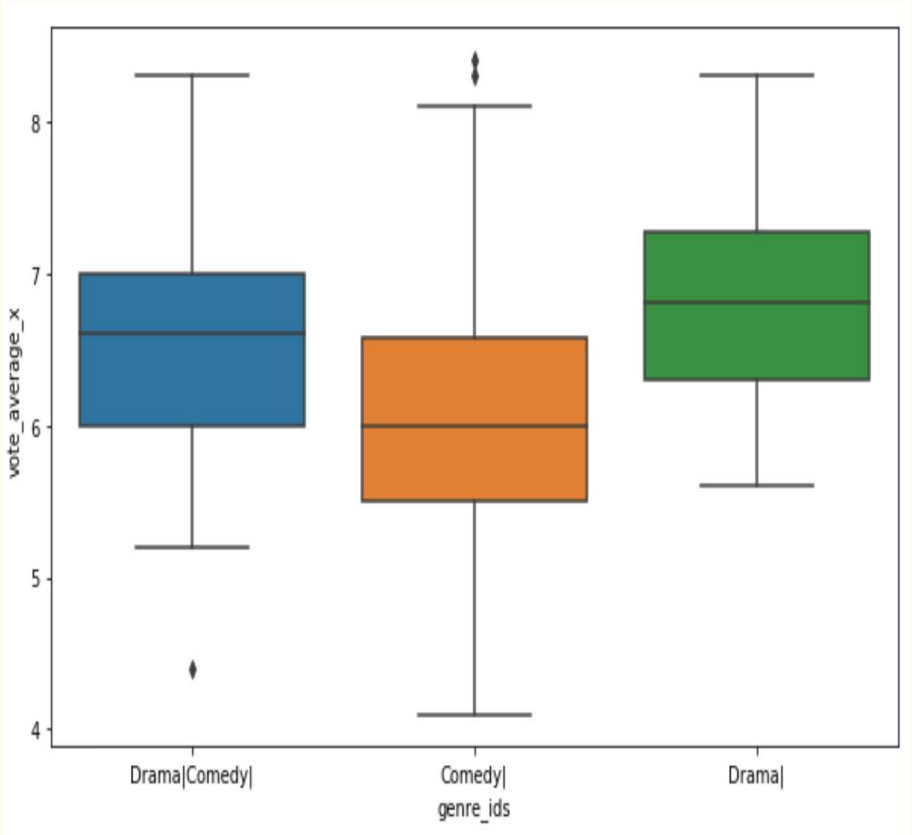


Genre - EDA (cont'd)

...decided to further focus on those genres that match our in-house expertise:

1. Drama
2. Comedy
3. Dramadey (Drama|Comedy)

Indication of a difference between the three genres



Genre - Hypothesis

H0 Null Hypothesis:

There is no significant difference in user rating between movie genres

Mean Rating of Drama|Comedy =
Mean Rating of Drama =
Mean Rating of Comedy

HA Alternative Hypothesis

There is a significant difference in user rating between movie genres

Mean Rating of Drama >
Mean Rating of Comedy <
Mean Rating of Drama|Comedy

Genre - Findings

Anova results: pr(>F) = 5.830832e-12

H0: Rejected

Post-hoc results (Tukey):

1. Comedy differed significantly to both Drama and Dramady ($p < 0.05$)
2. Drama and Dramedy exhibited no difference

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Comedy	Drama	0.7856	0.001	0.5354	1.0357	True
Comedy	Drama|Comedy	0.4994	0.001	0.2101	0.7888	True
Drama	Drama|Comedy	-0.2861	0.0734	-0.593	0.0207	False

null hypothesis between Comedy and Drama rejected

null hypothesis between Comedy and Drama|Comedy rejected

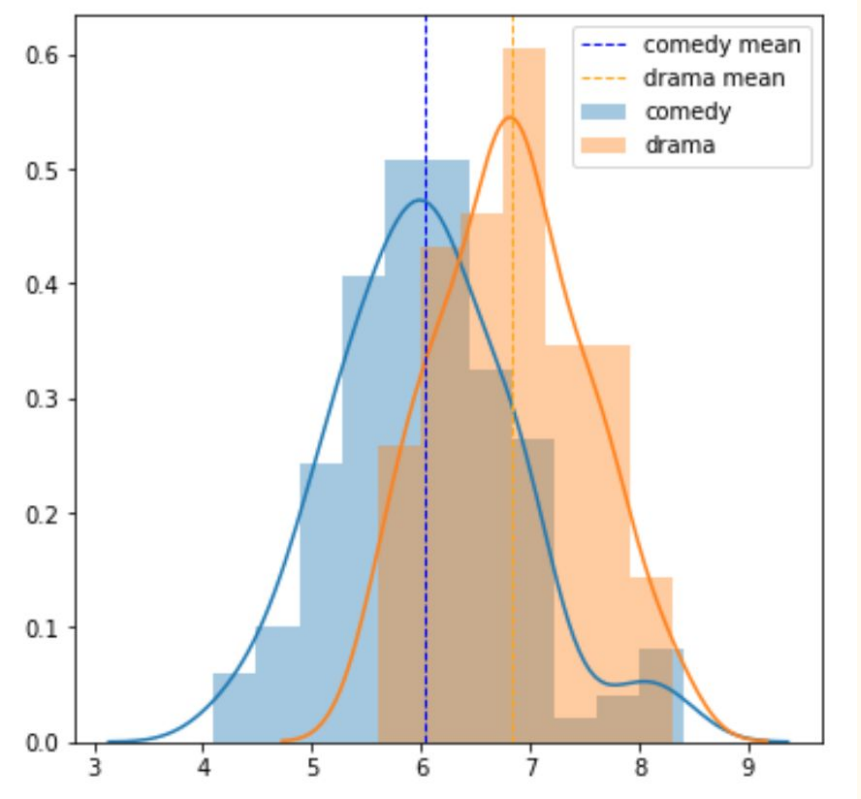
null hypothesis between Drama and Drama|Comedy **NOT** rejected

Genre - Comedy vs Drama

Sample size Comedy = **126**, Sample size Drama = **90**

Cohen's d = **-1.02,**

Power = **1.0**



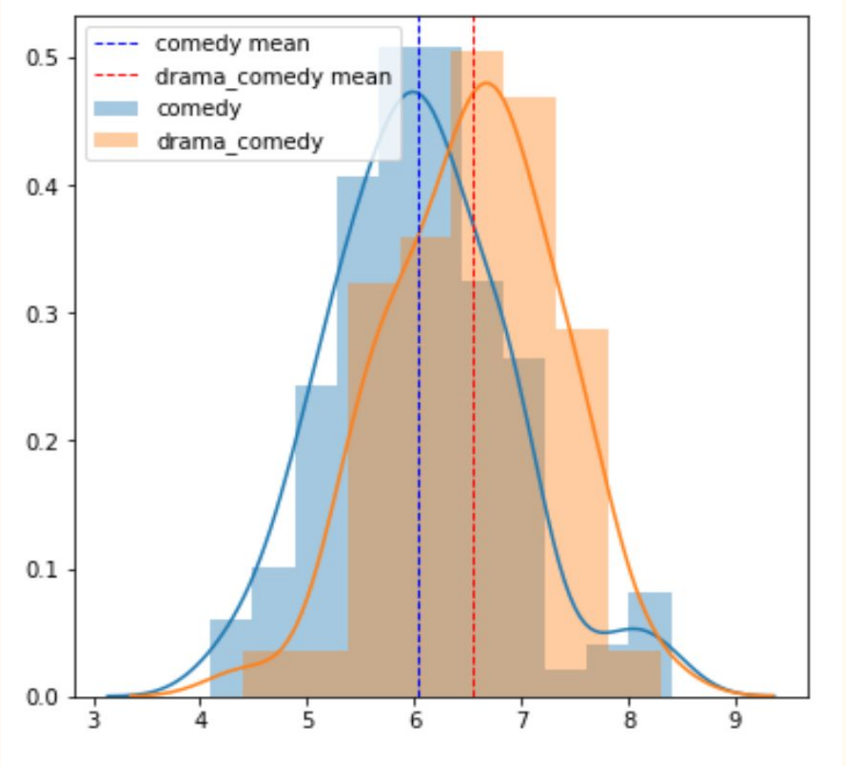
Genre - Comedy vs Dramadey

Sample size Comedy = **126**, Sample size Dramadey = **57**

Cohen's d = **-0.62,**

Power = **0.99**

Recommendation: Stick to pure Drama and avoid Comedy.



Hypothesis

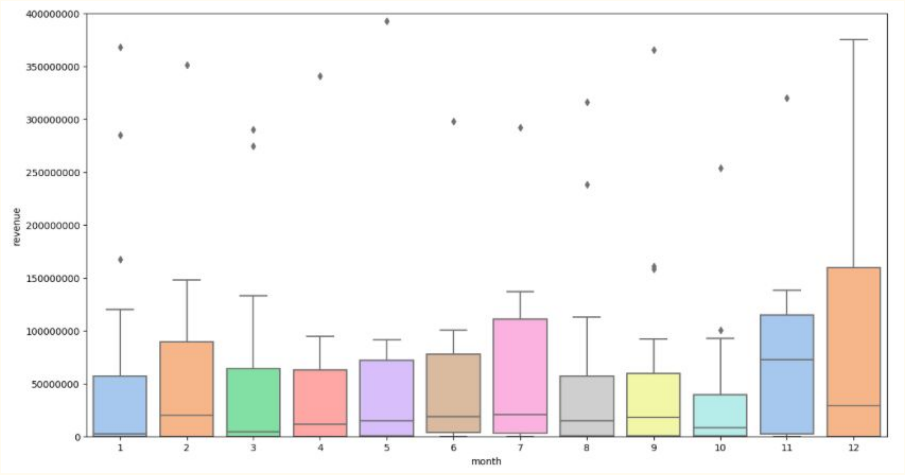
H0 Null Hypothesis:

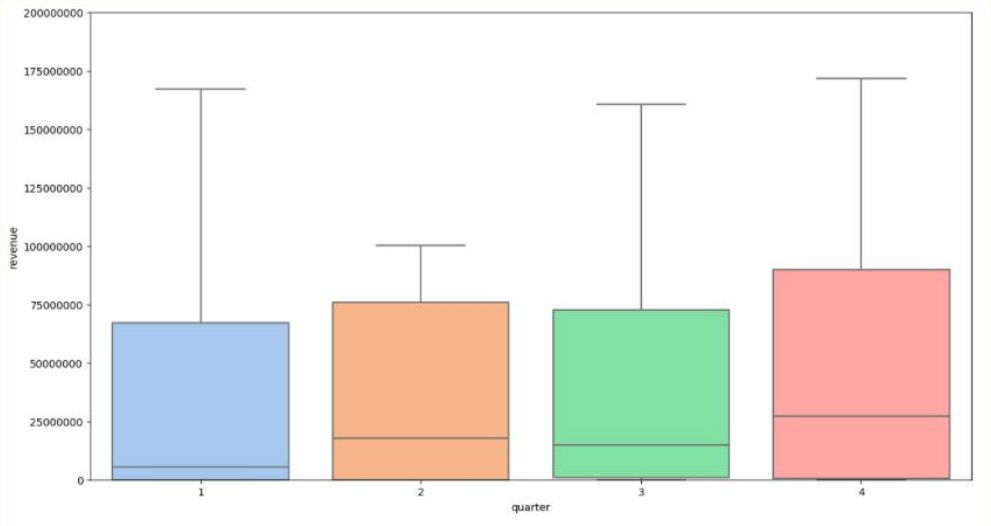
There is no significant difference in revenue for movies released at different times

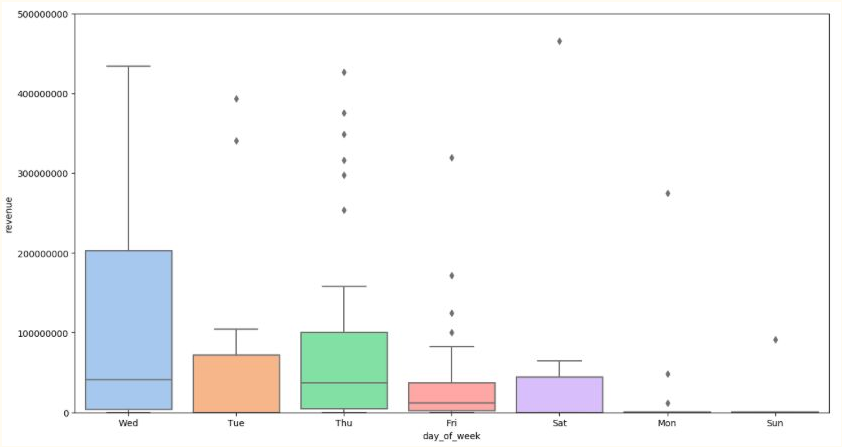
HA Alternative Hypothesis

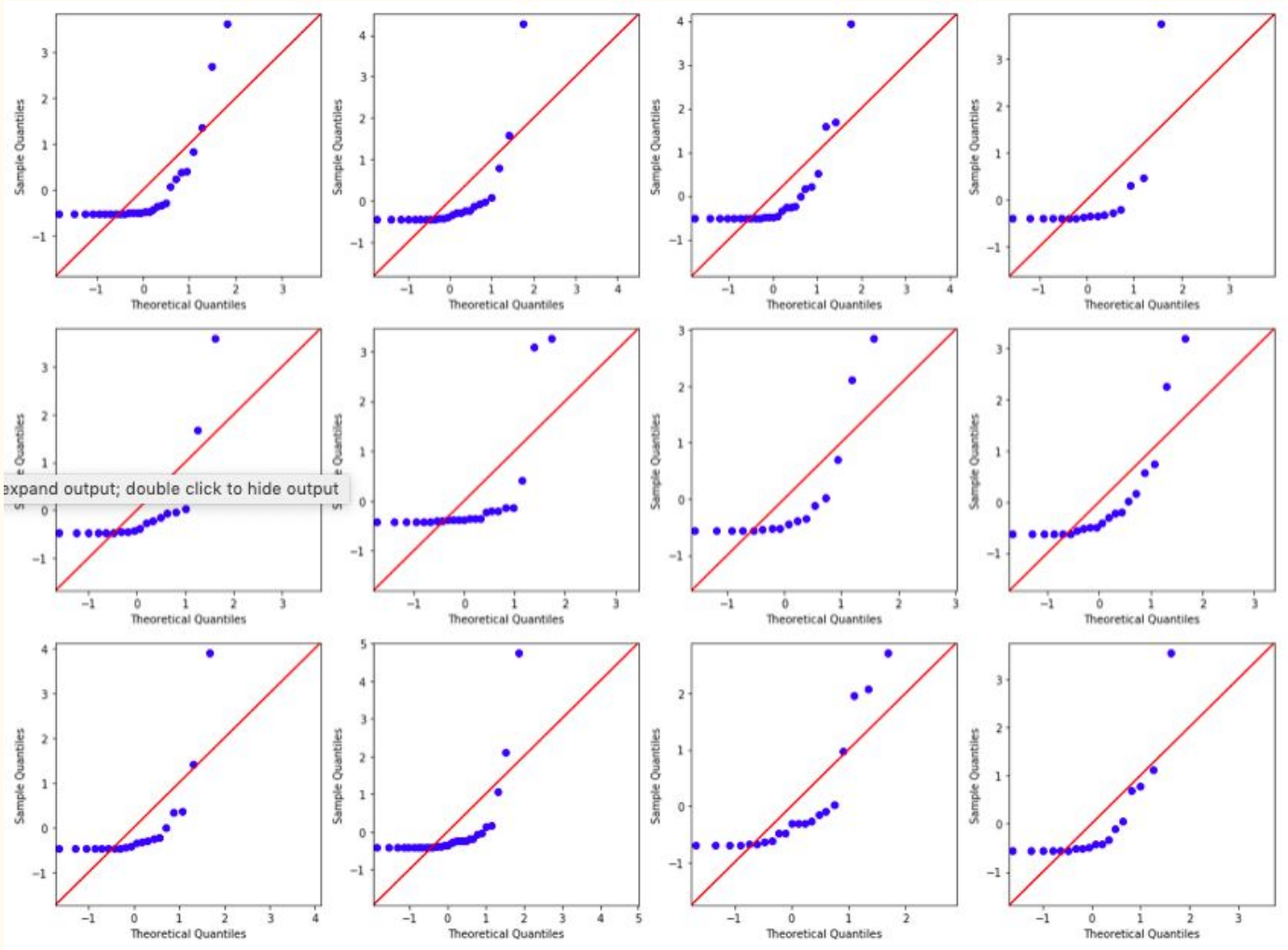
There is a significant difference in revenue for movies released at different times

Box and whisker plots of revenue distribution for: month, day and quarter of released (clockwise)









Residuals not normally distributed

Sample size too small to use central limit theorem

Decided to use non-parametric tests to assess null hypothesis

MannWhitneyU in place of T-Test

Kruskal-Wallis Test in place of ANOVA

Dunn Test in place of Tukey

```
1 kw_test(data1,data2,data3,data4,data5,data6,data7,data8,data9,data10,data11,data12)
```

```
Statistics=9.109, p=0.612
Same distributions (fail to reject H0)
```

```
1 kw_test(data1,data2,data3,data4)
```

```
Statistics=3.710, p=0.294
Same distributions (fail to reject H0)
```

```
1 kw_test(data_m,data_t,data_w,data_th,data_f,data_s,data_su)
```

```
Statistics=32.482, p=0.000
Different distributions (reject H0)
```

Kruskal-Wallis test results for month, quarter and day of release (from top to bottom).

Null hypothesis not rejected for quarter or month of release

But null hypothesis is rejected for day of week

```
kw_test(data_m,data_t,data_w,data_th,data_f,data_s,data_su)
```

```
Statistics=32.482, p=0.000
Different distributions (reject H0)
```

```
data_list = [data_m,data_t,data_w,data_th,data_f,data_s,data_su]

dunn_test(data_list)
```

	1	2	3	4	5	6	7
1	-1.000000	1.000000	0.002807	0.012901	0.252334	1.000000	1.000000
2	1.000000	-1.000000	0.130513	0.382736	1.000000	1.000000	0.889961
3	0.002807	0.130513	-1.000000	1.000000	0.110578	0.145988	0.007297
4	0.012901	0.382736	1.000000	-1.000000	0.477059	0.349721	0.024061
5	0.252334	1.000000	0.110578	0.477059	-1.000000	1.000000	0.252334
6	1.000000	1.000000	0.145988	0.349721	1.000000	-1.000000	1.000000
7	1.000000	0.889961	0.007297	0.024061	0.252334	1.000000	-1.000000

```
x=df_revenue['revenue'][df_revenue['day_of_week'] == 'Wed']
y=df_revenue['revenue'][df_revenue['day_of_week'] != 'Wed']
```

```
mann_whitney_u(x,y)
```

```
Statistics=4057.500, p=0.001
Different distribution (reject H0)
```

```
# estimation of standardised effect score
4057.5/(len(x)*len(y))
```

```
0.3598669623059867
```

```
stat, p = mannwhitneyu(x, y, alternative='greater')
print('Statistics=%.3f, p=%.3f' % (stat, p))
```

```
Statistics=7217.500, p=0.001
```

```
# estimation of standardised effect score if you assum one sided mann whitney test
stat/(len(x)*len(y))
```

```
0.6401330376940133
```

Ran a Dunn Test on day of release

Found several pairwise significant results

To find best day to release film, did a Mann Whitney U Test for Wednesday against not Wednesday. Found a significant result for this.

Used a crude estimator for the effect size to get an effect size of 0.64