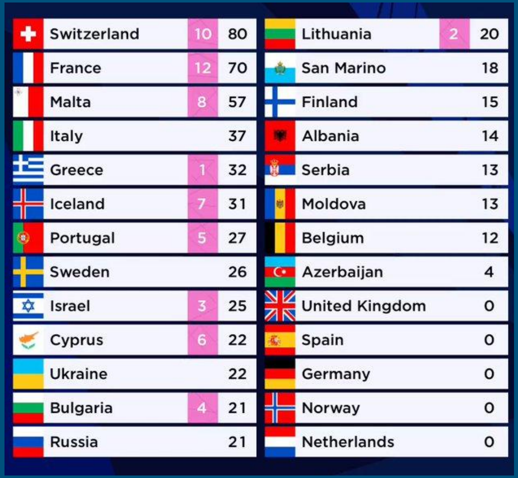
Eurovision Voting Trends

Jake Simon
22 September 2021

So Many Songs, So Few to Vote For!

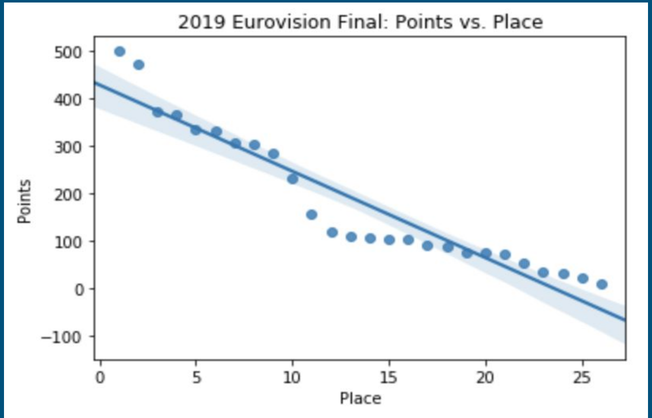
- The Eurovision Song Contest has a 2-tier voting system: jury and public televote
  - In each, voters pick their ten favorite songs, and assign 12 to their favorite, 10 to their 2nd, 8 to their 3rd, and 7-1 points for their 4th-10th favorites.
- What do we want to know?
  - Does the running order affect a song's point results?
  - As the contest has grown, more entries have been added to the contest each year.
    - As of 2021, there are 26 finalists, and 39 entries total!
    - Has this growth swayed the voters in any direction? Is it too much for them?



Switzerland	10	80
France	12	70
Malta	8	57
Italy		37
Greece	1	32
Iceland	7	31
Portugal	5	27
Sweden		26
Israel	3	25
Cyprus	6	22
Ukraine		22
Bulgaria	4	21
Russia		21
Lithuania	2	20
San Marino		18
Finland		15
Albania		14
Serbia		13
Moldova		13
Belgium		12
Azerbaijan		4
United Kingdom		0
Spain		0
Germany		0
Norway		0
Netherlands		0

The Tools

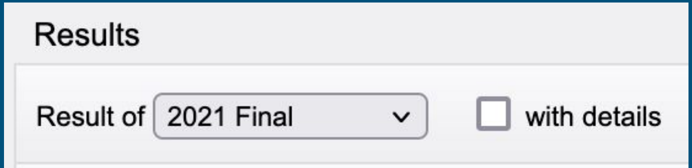
- How can we find this out?
  - We will use a linear regression model to detect positive or negative correlation as the order increases.



- What will we use to fulfill this task?
  - R and RStudio
  - SQL (SQLite package via RStudio)
  - Python and its numerous statistical packages
    - I.e. Pandas, Numpy, Matplotlib, etc.

Let’s Collect the Data

- Using eschome.net, we will download the “Results” of several Eurovision finals, starting with 2021.
- After that, we will save the page as “HTML only” to our “html database”.
- Once that’s there, we can now convert the data, using R and SQL!



2021, Final, Netherlands, Rotterdam, Rotterdam Ahoy, 22.5.2021

Place	Points	No.	Country	Performer
1	524	24	Italy	Måneskin
2	499	20	France	Barbara Pravi
3	432	11	Switzerland	Gjon's Tears
4	378	12	Iceland	Daði og Gagnamagnið
5	364	19	Ukraine	Go_A
6	301	16	Finland	Blind Channel

R and SQL: Converting and Cleaning the Data

```
```{r libraries}
# Libraries for HTML reading in R
library(xml2)
library(rvest)

# Libraries for SQL
library(RSQLite)
library(sqldf)
library(DBI)
```
```

The first step is to get the proper libraries. Once those are set, we can move on to the following steps:

```
```{r set_year}
year <- 2021
```
```

Set the year.

```
```{r read_html_fun}
# Reads the html file into the R program, and splits the table from the other data:
esc <- read_html(paste0("html database/Eurovision Song Contest Database ",
                 year, ".html")) %>% html_table()

# The table is the second element from the html_table function.
esc_table <- esc[[2]]

esc_finals <- read_html(paste0("html finals database/Eurovision Song Contest Database ",
                 year, ".html")) %>% html_table()

esc_finals_table <- esc_finals[[2]]

# Remove the empty columns that are there for some reason:
esc_table <- esc_table[,-1]
esc_finals_table <- esc_finals_table[,-4]

# We also need to rename the Order column "No." to "Song_Order", to avoid syntax errors:
colnames(esc_finals_table)[3] = "Song_Order"
```
```

Convert the html files to tables, and clean data.

```
```{r sql_skills}
# This connects us to the SQLite file for our dataset:
esc_db <- dbConnect(RSQLite::SQLite(), "esc-db.sqlite")

# Now, we can take that table, and add it to the SQL database:
dbWriteTable(esc_db, "ESC", esc_table, overwrite=T)
dbWriteTable(esc_db, "ESC_FINALS", esc_finals_table, overwrite=T)

# Let's add a new column to the table called "Points_Earned".
dbSendQuery(esc_db, '
            ALTER TABLE ESC_FINALS
            ADD COLUMN Percent_Earned DOUBLE')

# We'll take our new column, and calculate the percentage of points earned by country.
# Note: The denominator, 24 * (# total entries-1), is the max. points a country could
score.
dbSendStatement(esc_db, '
            UPDATE ESC_FINALS
            SET Percent_Earned =
              (Points+0.0) / (12 * 2.0 * ((SELECT COUNT(*) FROM ESC)-1))')

# Now, let's take the data and subset it by the following columns:
le_table <- dbGetQuery(esc_db, '
              SELECT Place, Points, Percent_Earned, Song_Order, Country,
              (SELECT COUNT(*) FROM ESC) AS Participants
              FROM ESC_FINALS')

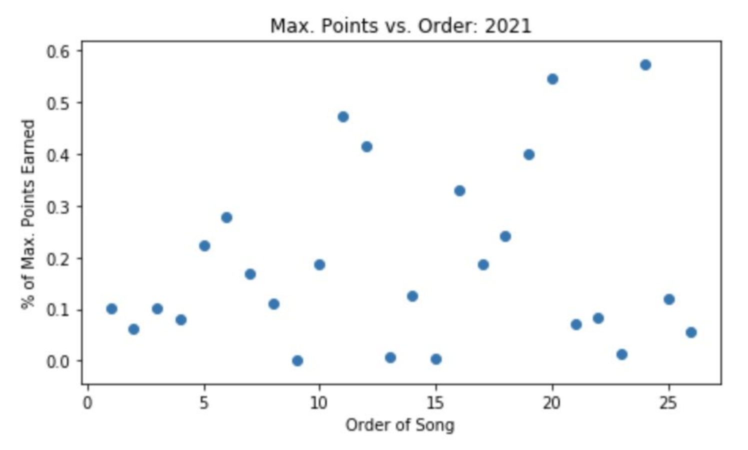
# Since this is R, we can assign the SQL table to an R table, and export that as a
# .csv file for our Python project:
write.csv(le_table, file=paste0("csv database/", year, " Eurovision Final Results.csv"))

# Now, let's disconnect from the SQLite server, and head on over to Python!
dbDisconnect(esc_db)
```
```

Use SQL to select and modify columns, then use R again to write said columns into .csv file.

Python: What Does It Look Like?

- Here is the head of the 2021 dataset, along with a scatter plot, comparing the final entrants' points to their song order.
  - Pandas was used to create a DataFrame.
  - Matplotlib was used to create the graph.

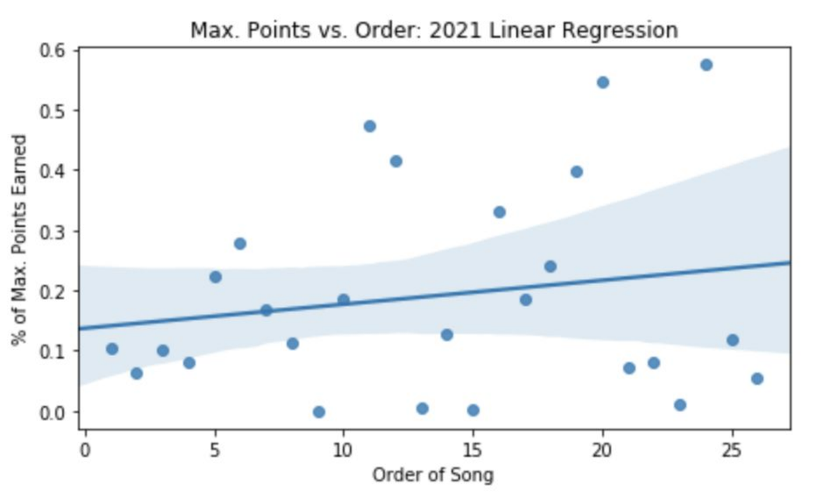


	Year	Place	Points	Percent_Earned	Song_Order	Country	Participants
0	2021	16	94	0.103070	1	Cyprus	39
1	2021	21	57	0.062500	2	Albania	39
2	2021	17	93	0.101974	3	Israel	39
3	2021	19	74	0.081140	4	Belgium	39
4	2021	9	204	0.223684	5	Russia	39

Regression Model: Does Song Order Affect Points?

- Let's see if the song order has any significant correlation with percentage of points earned:
    - The 'regplot' function is used to graph the data, while 'ols.fit()' is used to create the summary.

```
Eurovision 2021
                            OLS Regression Results
==============================================================================
Dep. Variable:         Percent_Earned   R-squared:                       0.032
Model:                            OLS   Adj. R-squared:                 -0.008
Method:                 Least Squares   F-statistic:                    0.7938
Date:                Tue, 21 Sep 2021   Prob (F-statistic):              0.382
Time:                        22:05:01   Log-Likelihood:                 10.143
No. Observations:                  26   AIC:                            -16.29
Df Residuals:                      24   BIC:                            -13.77
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1372      0.069      1.992      0.058      -0.005      0.279
Song_Order     0.0040      0.004      0.891      0.382      -0.005      0.013
==============================================================================
Omnibus:                        2.815   Durbin-Watson:                   1.876
Prob(Omnibus):                  0.245   Jarque-Bera (JB):                2.457
Skew:                           0.697   Prob(JB):                        0.293
Kurtosis:                       2.428   Cond. No.                         31.9
==============================================================================
```



Turns out there is hardly any *linear* correlation, but there is still an interesting pattern as song order increases. It's *diverging*!

How to Make Sense of the Diverging Pattern

- It appears to have a funnel-like shape, as the Order of Song increases.
- How can we analyze this trend via regression?
  - Take the fitted model from before, and turn the point residuals as the new response!
  - Rather than a direct positive / negative correlation, we can now see if the points become less predictable over the course of the contest! (in other words, as the Order of Song increases)


```
ols_lm = ols('Percent_Earned ~ Song_Order', data=test).fit()

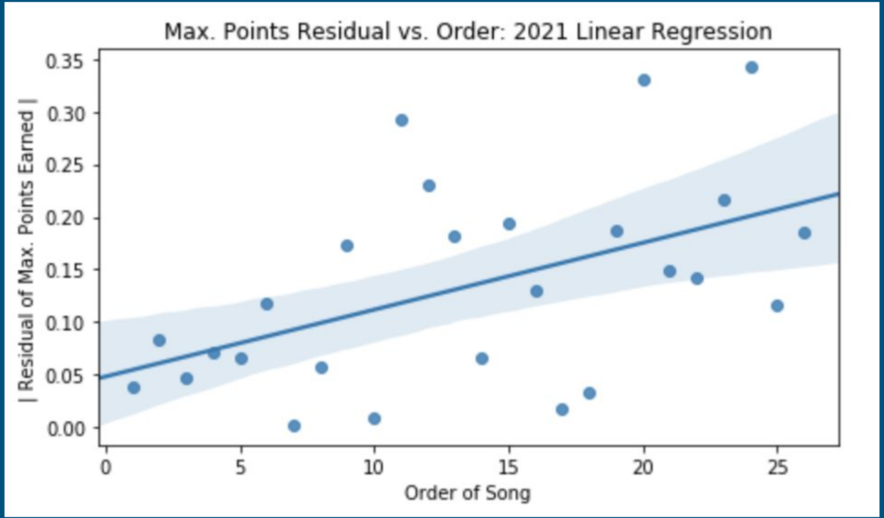
test['Perc_Mag_Res'] = abs(ols_lm.resid)
```

We will take the residual’s absolute value to measure predictability. Lower value = more predictable

Residual Regression: How Does Song Order Affect Predictability?

- Let's take a look at the scatter plot, adjusted for residual magnitude.
  - This seems a lot closer to a significant correlation!
- How does this impact the regression model's fit?

```
Eurovision 2021
                            OLS Regression Results
==============================================================================
Dep. Variable:            Perc_Mag_Res   R-squared:                       0.256
Model:                             OLS   Adj. R-squared:                  0.225
Method:                  Least Squares   F-statistic:                     8.249
Date:                 Tue, 21 Sep 2021   Prob (F-statistic):            0.00839
Time:                         22:34:55   Log-Likelihood:                 28.333
No. Observations:                   26   AIC:                            -52.67
Df Residuals:                       24   BIC:                            -50.15
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0480      0.034      1.405      0.173      -0.023       0.119
Song_Order     0.0064      0.002      2.872      0.008       0.002       0.011
==============================================================================
Omnibus:                        1.001   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.606   Jarque-Bera (JB):                0.887
Skew:                           0.412   Prob(JB):                        0.642
Kurtosis:                       2.626   Cond. No.                         31.9
==============================================================================
```



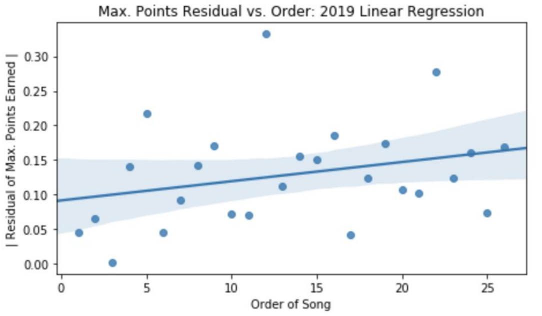
Quite a lot actually! We can say with over 99% confidence, the further a song is in the contest (higher order), the less predictable its points become!

What does this finding mean?

- Since the original regression model gets less accurate as song order increases, this could imply voters are a lot more fickle or judgmental towards later songs when they vote.
    - People could also tune out until the show gets closer to the results, causing them to pay more attention to the latter songs.
    - On the other hand, it could mean voters get disinterested as the contest goes along, and put less thought into who they vote for.
        - This seems less likely as it would imply all points would go down towards the end, which they did not, according to our first graph.

Was this a fluke one year?

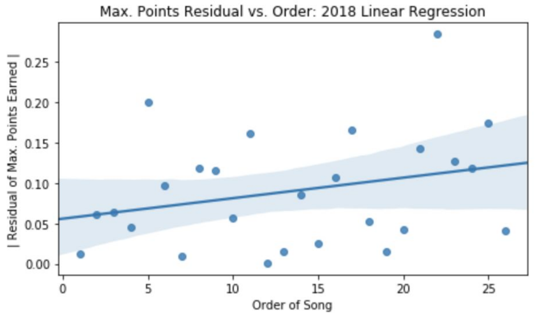
Let’s take the previous three years, and measure their residuals.



Eurovision 2019

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Perc_Mag_Res   R-squared:                       0.083
Model:                            OLS   Adj. R-squared:                  0.045
Method:                 Least Squares   F-statistic:                     2.166
Date:                Tue, 21 Sep 2021   Prob (F-statistic):              0.154
Time:                        23:11:18   Log-Likelihood:                 32.428
No. Observations:                  26   AIC:                            -60.86
Df Residuals:                      24   BIC:                            -58.34
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0912      0.029      3.119      0.005       0.031       0.151
Song_Order     0.0028      0.002      1.472      0.154      -0.001       0.007
==============================================================================
Omnibus:                        8.685   Durbin-Watson:                   2.400
Prob(Omnibus):                  0.013   Jarque-Bera (JB):                6.697
Skew:                           1.079   Prob(JB):                       0.0351
Kurtosis:                       4.236   Cond. No.                         31.9
==============================================================================
```

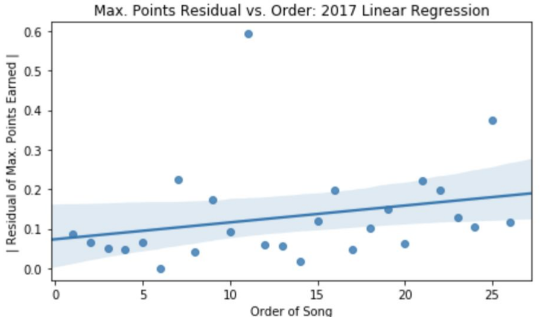
2019: p-value = 0.154



Eurovision 2018

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Perc_Mag_Res   R-squared:                       0.078
Model:                            OLS   Adj. R-squared:                  0.040
Method:                 Least Squares   F-statistic:                     2.033
Date:                Tue, 21 Sep 2021   Prob (F-statistic):              0.167
Time:                        23:14:20   Log-Likelihood:                 34.030
No. Observations:                  26   AIC:                            -64.06
Df Residuals:                      24   BIC:                            -61.54
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0558      0.027      2.031      0.054      -0.001       0.112
Song_Order     0.0025      0.002      1.426      0.167      -0.001       0.006
==============================================================================
Omnibus:                        3.358   Durbin-Watson:                   1.856
Prob(Omnibus):                  0.187   Jarque-Bera (JB):                2.236
Skew:                           0.714   Prob(JB):                        0.327
Kurtosis:                       3.163   Cond. No.                         31.9
==============================================================================
```

2018: p-value = 0.167



Eurovision 2017

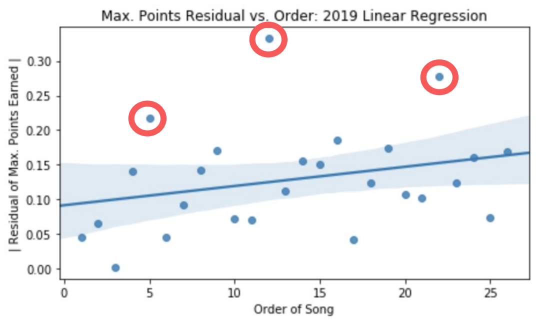
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Perc_Mag_Res   R-squared:                       0.068
Model:                            OLS   Adj. R-squared:                  0.030
Method:                 Least Squares   F-statistic:                     1.762
Date:                Tue, 21 Sep 2021   Prob (F-statistic):              0.197
Time:                        23:17:08   Log-Likelihood:                 18.700
No. Observations:                  26   AIC:                            -33.40
Df Residuals:                      24   BIC:                            -30.88
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0735      0.050      1.483      0.151      -0.029       0.176
Song_Order     0.0043      0.003      1.327      0.197      -0.002       0.011
==============================================================================
Omnibus:                       34.184   Durbin-Watson:                   2.402
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               87.480
Skew:                           2.561   Prob(JB):                     1.01e-19
Kurtosis:                      10.383   Cond. No.                         31.9
==============================================================================
```

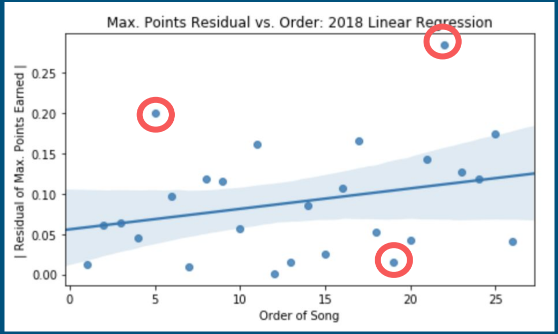
2017: p-value = 0.197

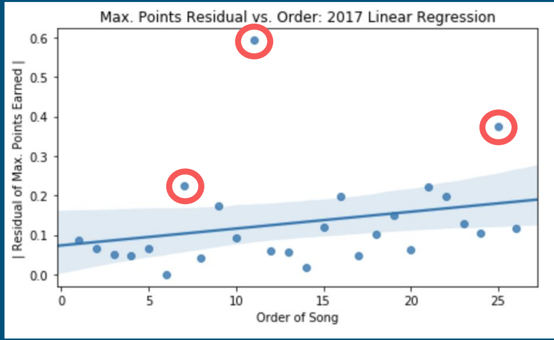
Maybe? Let’s remove the top 3 outliers per year.

Taking out the 3 songs furthest from the regression line removes any outstanding influence from singular points.

Although most points are within range of a positive correlation, the least accurate points (usually the songs with the most points) are in their own world mathematically and visually.







Song_Order	Country	Place	Points	Percent_Earned	Perc_Mag_Res
12	Netherlands	1	498	0.518750	0.332306
22	Italy	2	472	0.491667	0.277886
5	Russia	3	370	0.385417	0.218108

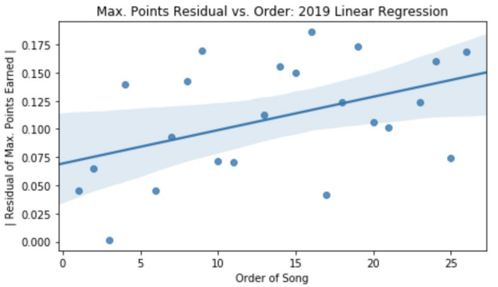
Song_Order	Country	Place	Points	Percent_Earned	Perc_Mag_Res
22	Israel	1	529	0.524802	0.283946
5	Austria	3	342	0.339286	0.199494
19	Moldova	10	209	0.207341	0.015679

Song_Order	Country	Place	Points	Percent_Earned	Perc_Mag_Res
11	Portugal	1	758	0.770325	0.592798
25	Bulgaria	2	615	0.625000	0.375211
7	Moldova	3	374	0.380081	0.223200

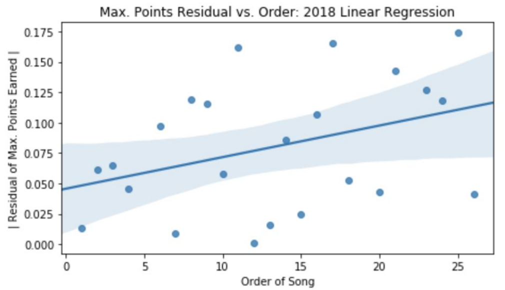
Now, let’s regraph the residual regression models! This should give us a more accurate reading of the data and its trends.

Now what do our regressions look like?

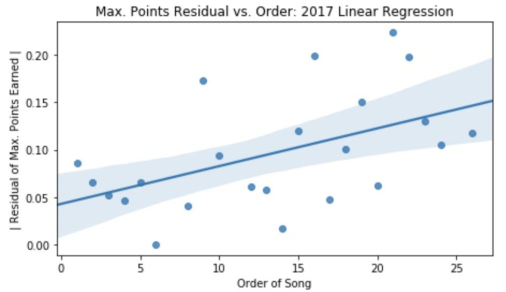
These are the regression models without their top 3 outliers.




```
Eurovision 2019
                            OLS Regression Results
==============================================================================
Dep. Variable:            Perc_Mag_Res   R-squared:                       0.203
Model:                             OLS   Adj. R-squared:                  0.165
Method:                  Least Squares   F-statistic:                     5.357
Date:                 Tue, 21 Sep 2021   Prob (F-statistic):             0.0308
Time:                         23:46:21   Log-Likelihood:                 39.016
No. Observations:                   23   AIC:                            -74.03
Df Residuals:                       21   BIC:                            -71.76
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0693      0.020      3.487      0.002       0.028       0.111
Song_Order     0.0030      0.001      2.315      0.031       0.000       0.006
==============================================================================
Omnibus:                        1.314   Durbin-Watson:                   2.317
Prob(Omnibus):                  0.518   Jarque-Bera (JB):                0.888
Skew:                          -0.068   Prob(JB):                        0.642
Kurtosis:                       2.047   Cond. No.                         32.0
==============================================================================
```




```
Eurovision 2018
                            OLS Regression Results
==============================================================================
Dep. Variable:            Perc_Mag_Res   R-squared:                       0.138
Model:                             OLS   Adj. R-squared:                  0.097
Method:                  Least Squares   F-statistic:                     3.360
Date:                 Tue, 21 Sep 2021   Prob (F-statistic):             0.0810
Time:                         23:49:17   Log-Likelihood:                 36.903
No. Observations:                   23   AIC:                            -69.81
Df Residuals:                       21   BIC:                            -67.54
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0456      0.022      2.110      0.047       0.001       0.091
Song_Order     0.0026      0.001      1.833      0.081      -0.000       0.006
==============================================================================
Omnibus:                        3.022   Durbin-Watson:                   2.358
Prob(Omnibus):                  0.221   Jarque-Bera (JB):                1.284
Skew:                           0.011   Prob(JB):                        0.526
Kurtosis:                       1.843   Cond. No.                         31.1
==============================================================================
```




```
Eurovision 2017
                            OLS Regression Results
==============================================================================
Dep. Variable:            Perc_Mag_Res   R-squared:                       0.257
Model:                             OLS   Adj. R-squared:                  0.222
Method:                  Least Squares   F-statistic:                     7.261
Date:                 Tue, 21 Sep 2021   Prob (F-statistic):             0.0136
Time:                         23:52:48   Log-Likelihood:                 36.092
No. Observations:                   23   AIC:                            -68.18
Df Residuals:                       21   BIC:                            -65.91
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0431      0.023      1.908      0.070      -0.004       0.090
Song_Order     0.0040      0.001      2.695      0.014       0.001       0.007
==============================================================================
Omnibus:                        1.371   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.504   Jarque-Bera (JB):                1.244
Skew:                           0.483   Prob(JB):                        0.537
Kurtosis:                       2.397   Cond. No.                         31.6
==============================================================================
```

2019: p-value = 0.031

2018: p-value = 0.081

2017: p-value = 0.014

Conclusion

- Based on the previous three contests, 2021's measurement of unpredictability seems to indicate a trend.
    - We can say with at least 90% confidence (96.9% excluding 2018) that later songs tend to vary greater and less predictably than earlier songs.
- An interesting thing to note is that all four contests have had around 40 total participants each, and 26 final participants each.
    - Based on this logic of prediction, voters tend to be more sporadic with their voting later in the contest.
        - This could be attributed to later songs being more fresh in the voter's head when recalling all the songs.
        - In the future, we can test this logic on older contests with fewer entries to see if the residuals and general distribution are less noticeable.

Links to My Programs and Repository

- GitHub:
  - Project Repository: https://github.com/jakesimon2/ESC-Voting-Trends
    - This site contains all my code, data sources, and presentation slides.
- Eurovision Song Picker (also on GitHub, created in 2020)
  - https://github.com/jakesimon2/Eurovision-Song-Picker
- Thanks to eschome for keeping great track of all Eurovision Song Contests!
  - Link to Website, where I downloaded the data from: https://www.eschome.net/index.html