۱ تجزیه و تحلیل عوامل

۱-۱ زیربخش ۱

فرض صفر و فرض جایگزین:

-فرض صفر، این را بیان می‌کند که میانگین امتیازهای محبوبیت در چهار سبک موسیقی یکسان است. و تفاوت significant ای (قابل توجهی) میان این میانگین‌ها وجود ندارد.

$$H_0 : \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 \tag{۱}$$

-فرض جایگزین، این را بیان می‌کند که میانگین امتیازهای محبوبیت در حداقل یکی از چهار سبک موسیقی، با بقیه برابر نیست. و تفاوت significant ای (قابل توجهی) با بقیه‌ی میانگین‌ها دارد.

$$\begin{aligned} H_A : \quad & \mu_1 \neq \mu_2 \quad or \quad \mu_1 \neq \mu_3 \\ or \quad & \mu_1 \neq \mu_4 \quad or \quad \mu_2 \neq \mu_3 \\ or \quad & \mu_2 \neq \mu_4 \quad or \quad \mu_3 \neq \mu_4 \end{aligned} \tag{۲}$$

۱-۲ زیربخش ۲

برای انجام آزمون ANOVA باید ابتدا یک سری پارامتر محاسبه شود:

میانگین امتیازهای محبوبیت مربوط به سبک Pop:

$$\bar{y}_{Pop} = \frac{85 + 88 + 78 + 80}{4} = \frac{331}{4} = 82.75 \tag{۳}$$

میانگین امتیازهای محبوبیت مربوط به سبک Hip-Hop:

$$\bar{y}_{Hip-Hop} = \frac{75+72+80+77}{4} = \frac{304}{4} = 76 \quad (۴)$$

میانگین امتیازهای محبوبیت مربوط به سبک Rock:

$$\bar{y}_{Rock} = \frac{23+11+31+8}{4} = \frac{73}{4} = 18.25 \quad (۵)$$

میانگین امتیازهای محبوبیت مربوط به سبک Electronic:

$$\bar{y}_{Electronic} = \frac{80+82+76+75}{4} = \frac{313}{4} = 78.25 \quad (۶)$$

میانگین امتیازهای محبوبیت مربوط به همه‌ی سبک‌ها:

$$\bar{y} = \frac{331+304+73+313}{16} = \frac{1021}{16} = 63.8125 \quad (۷)$$

Sum of Squares Total که تنوع کلی داده‌ها را اندازه‌گیری می‌کند:

$$\begin{aligned} SST &= \sum_{i=1}^{18}(y_i - \bar{y})^2 \\ &= (85-63.8125)^2 + (88-63.8125)^2 + (78-63.8125)^2 + (80-63.8125)^2 \\ &+(75-63.8125)^2 + (72-63.8125)^2 + (80-63.8125)^2 + (77-63.8125)^2 \\ &+(23-63.8125)^2 + (11-63.8125)^2 + (31-63.8125)^2 + (8-63.8125)^2 \\ &+(80-63.8125)^2 + (82-63.8125)^2 + (76-63.8125)^2 + (75-63.8125)^2 \\ &= (21.1875)^2 + (24.1875)^2 + (14.1875)^2 + (16.1875)^2 + (11.1875)^2 \\ &+(8.1875)^2 + (16.1875)^2 + (13.1875)^2 + (-40.8125)^2 + (-52.8125)^2 \\ &+(-32.8125)^2 + (-55.8125)^2 + (16.1875)^2 + (18.1875)^2 + (12.1875)^2 \\ &+(11.1875)^2 = 448.91 + 585.04 + 201.29 + 262.04 + 125.16 \\ &+67.04 + 262.04 + 173.91 + 1665.66 + 2789.16 + 1076.66 \\ &+3115.04 + 262.04 + 330.79 + 148.54 + 125.16 = 11638.48 \end{aligned} \quad (۸)$$

Sum of Squares Groups که تنوع بین گروهی را اندازه‌گیری می‌کند:

$$
\begin{aligned}
SSG &= \sum_{j=1}^{3} n_j(\bar{y}_j - \bar{y})^2 \\
&= 4(82.75 - 63.8125)^2 + 4(76 - 63.8125)^2 \\
&+4(18.25 - 63.8125)^2 + 4(78.25 - 63.8125)^2 \\
&= 4(18.9375)^2 + 4(12.1875)^2 + 4(-45.5625)^2 + 4(14.4375)^2 \\
&= 4(358.63) + 4(148.54) + 4(2075.94) + 4(208.44) \\
&= 1434.52 + 594.16 + 8303.76 + 833.76 = 11166.2
\end{aligned}
\tag{۹}
$$

Sum of Squares Error:

$$SSE = SST - SSG = 11638.48 - 11166.2 = 472.28 \tag{۱۰}$$

درجه آزادی کل:

$$df_T = n - 1 = 16 - 1 = 15 \tag{۱۱}$$

درجه آزادی گروه:

$$df_G = K - 1 = 4 - 1 = 3 \tag{۱۲}$$

درجه آزادی خطا:

$$df_E = df_T - df_G = 15 - 3 = 12 \tag{۱۳}$$

mean squares گروه:

$$MSG = \frac{SSG}{df_G} = \frac{11166.2}{3} = 3722.07 \tag{۱۴}$$

mean squares خطا:

$$MSE = \frac{SSE}{df_E} = \frac{472.28}{12} = 39.36 \quad (۱۵)$$

مقدار آماره‌ی F:

$$F = \frac{MSG}{MSE} = \frac{3722.07}{39.36} = 94.56 \quad (۱۶)$$

جدول ANOVA به فرم زیر است:

		DF	Sum sq	Mean sq	F value	Pr (>F)
Group	Class	DFG=k-۱	SSG	MSG=SSG/DFG	MSG/MSE	؟
Error	Residuals	DFE=DFT-DFG	SSE	MSE=SSE/DFE		
	Total	DFT=n-۱	SST			

با جایگذاری به فرم زیر حاصل می‌شود:

		DF	Sum sq	Mean sq	F value	Pr >F)
Group	Class	۳	۱۱۱۶۶.۲	۳۷۲۲.۰۷	۹۴.۵۶	<۰.۰۰۱
Error	Residuals	۱۲	۴۷۲.۲۸	۳۹.۳۶		
	Total	۱۵	۱۱۶۳۸.۴۸			

به کمک یک ماشین حساب آنلاین مربوط به آماره‌ی F، برای توزیع F با درجه آزادی‌های ۳ و ۱۲، و مقدار آماره برابر با ۹۴.۵۶ مقدار p-value بسیار کوچک بدست آمد. با توجه به این‌که این مقدار، از سطح significance که برابر ۰.۰۵ است، کمتر است، فرض صفر رد می‌شود. و نمی‌توان ادعا کرد که هیچ تفاوت significant ای بین میانگین امتیازهای محبوبیت در چهار سبک موسیقی وجود ندارد و این میانگین‌ها با هم برابرند. و احتمالاً حداقل یکی از این میانگین‌ها، نسبت به بقیه، دارای تفاوت قابل توجهی‌ست.

۳-۱ زیربخش ۳

(برای حل این بخش، از کد R استفاده شده‌است.)

آزمون Post-hoc یک تست آماری‌ست که پس از یافتن تفاوت significant در یک آزمون ANOVA (تجزیه و تحلیل واریانس) یا آزمون مشابه، از آن استفاده می‌شود. و هدف آن تعیین گروه‌های خاصی است که از یکدیگر متمایز هستند و برای زمانی مفید

است که بیش از دو گروه مورد مقایسه قرار گرفته باشند. که با توجه به این‌که در این مسئله، ۴ گروه وجود دارد، از این منظر، استفاده از این آزمون مناسب است.
برای بررسی قابل توجه بودن تفاوت بین گروه‌ها از نظر میانگین هر یک از امتیازات نیز، از آزمون ANOVA استفاده می‌شود. (در سوال قبل، میانگین امتیازهای محبوبیت گروه‌ها، بررسی شده‌است و قابل توجه بودن میانگین گروه‌ها نشان داده شده‌است. اما این‌جا یک بار دیگر میانگین هر ۳ نوع امتیاز، با استفاده از آزمون ANOVA بررسی خواهند شد.)
مقدار p-value مربوط به آزمون ANOVA با استفاده از کد R برای هر یک از امتیازها محاسبه شد. مقدار p-value برای آزمون میانگین امتیازهای Popularity کم‌تر از مقدار ۰.۰۵ شد و این نشان‌دهنده‌ی وجود تفاوت قابل توجه میان میانگین امتیازات Popularity مربوط به ۴ سبک موسیقی است. مقدار p-value برای آزمون میانگین امتیازهای Danceability کم‌تر از مقدار ۰.۰۵ شد و این نشان‌دهنده‌ی وجود تفاوت قابل توجه میان میانگین امتیازات Danceability مربوط به ۴ سبک موسیقی است.
مقدار p-value برای آزمون میانگین امتیازهای Energy بیشتر از مقدار ۰.۰۵ شد و این نشان‌دهنده‌ی عدم وجود تفاوت قابل توجه میان میانگین امتیازات Energy مربوط به ۴ سبک موسیقی است.
با توجه به بررسی انجام شده برای قابل توجه بودن/نبودن تفاوت‌ها، آزمون Post-hoc روی دو امتیاز Popularity و Danceability قابل انجام است.

میانگین و انحراف معیار دو امتیاز Popularity و Danceability، برای هر یک از ۴ سبک موسیقی، به تفکیک، محاسبه شد:

genre	mean_popularity	sd_popularity	mean_danceability	sd_danceability
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Electronic	78.25	3.304038	8	0.8164966
Hip-Hop	76.00	3.366502	6	0.8164966
Pop	82.75	4.573474	7	0.8164966
Rock	18.25	10.688779	5	0.8164966

با توجه به این‌که تعداد گروه‌ها برابر ۴، درجه آزادی برابر ۱۲ است و سطح signifi-cance برابر ۰.۰۵ است، مقدار بحرانی روش Tukey's HSD برابر ۸.۲۱ بدست می‌آید.

با استفاده از مقدار بحرانی محاسبه شده، به کمک فرمول زیر، مقدار HSD برای هر یک از دو امتیاز Popularity و Danceability، بدست می‌آید:

$$HSD = CriticalValue * \sqrt{\frac{MSE}{n}} \quad (۱۷)$$

```
hsd for popularity =  12.87323
hsd for danceability =  1.675509
```

سپس تمام زوج سبک‌های ممکن در نظر گرفته می‌شود و اختلاف میانگین Popularity و Danceability آن‌ها دو به دو محاسبه می‌شود:

genre1	genre2	diff_mean_popularity	diff_mean_danceability
<chr>	<chr>	<dbl>	<dbl>
Electronic	Hip-Hop	2.25	2
Electronic	Pop	4.50	1
Electronic	Rock	60.00	3
Hip-Hop	Pop	6.75	1
Hip-Hop	Rock	57.75	1
Pop	Rock	64.50	2

سپس مقدار standard error مربوط به Popularity و Danceability با استفاده از فرمول زیر محاسبه شد:
(مقدار مربوط به Popularity برابر ۲.۲۲ و مقدار مربوط به Danceability برابر ۰.۲۹ بدست آمد.)

$$StandardErrorOfDifference = \sqrt{2 * \frac{MSE}{n}} \quad (۱۸)$$

در نهایت، با تقسیم اختلاف میانگین‌های هر یک از زوج سبک‌ها، بر standard error بدست آمده، مقدار t-value مربوط به Popularity و Danceability برای هر یک از زوج‌ها محاسبه شد:

genre1	genre2	diff_mean_popularity	diff_mean_danceability	t_value_popularity	t_value_danceability
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
Electronic	Hip-Hop	2.25	2	1.014454	6.928203
Electronic	Pop	4.50	1	2.028907	3.464102
Electronic	Rock	60.00	3	27.052094	10.392305
Hip-Hop	Pop	6.75	1	3.043361	3.464102
Hip-Hop	Rock	57.75	1	26.037640	3.464102
Pop	Rock	64.50	2	29.081001	6.928203

در آخر، برای بررسی این‌که کدام زوج گروه‌ها دارای تفاوت قابل توجه در میانگین مربوط به امتیازات خود هستند، مقدار t-value آن‌ها با مقدار HSD مقایسه شد. هر زوج گروه که مقدار t-value اختلاف یکی از امتیازاتشان بیشتر از مقدار HSD آن امتیاز باشد، از نظر میانگین آن امتیاز، دارای اختلاف قابل توجهی هستند. مطابق آنچه در شکل زیر نمایش داه شده است، تمامی زوج سبک‌های موسیقی از نظر میانگین Danceability دارای تفاوت قابل توجهی هستند. و از نظر میانگین Popularity سبک موسیقی Rock دارای تفاوت قابل توجه نسبت به بقیه‌ی سبک‌هاست:

genre1	genre2	diff_mean_popularity	diff_mean_danceability	significant_popularity_diff_mean	significant_danceability_diff_mean
<chr>	<chr>	<dbl>	<dbl>	<chr>	<chr>
Electronic	Hip-Hop	2.25	2	No	Yes
Electronic	Pop	4.50	1	No	Yes
Electronic	Rock	60.00	3	Yes	Yes
Hip-Hop	Pop	6.75	1	No	Yes
Hip-Hop	Rock	57.75	1	Yes	Yes
Pop	Rock	64.50	2	Yes	Yes

۴-۱ زیربخش ۴

(برای حل این بخش، از کد R استفاده شده‌است.)

می‌توان از ANOVA دو طرفه (two-way) استفاده کرد. در این نوع تست، تاثیر متقابل

دو عامل بر روی یک متغیر response برسی می‌شود.
پس یک تست ANOVA دو طرفه با در نظر گرفتن دو متغیر Danceability و Energy به عنوان متغیرهای مستقل و متغیر Popularity به عنوان متغیر پاسخ زده شد. دلیل انتخاب این دو متغیر به عنوان متغیرهای مستقل این است که آن‌ها به طور بالقوه می‌توانند بر امتیاز Popularity تأثیر بگذارند. با تجزیه و تحلیل اثر متقابل بین این دو متغیر و سبک موسیقی‌ها، می‌توانیم تعیین کنیم که آیا رابطه‌ی بین Danceability و Energy و امتیاز Popularity در سبک‌ها متفاوت است یا خیر.
به علاوه، این تست از این جهت موثر است، که هم تاثیر اصلی هر متغیر مستقل و هم تاثیر متقابل آن‌ها بر متغیر پاسخ را بررسی می‌کند. که این مسئله، درک جامع‌تری از روابط بین این متغیرها و تفاوت بین سبک‌ها ارائه می‌دهد. (البته اینجا تاثیر تمامی جایگشت‌های بقیه‌ی متغیرها بر متغیر Popularity، بررسی شد و آزمونی کلی‌تر انجام شد.)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
danceability	1	5089.50893	5089.508929	20358.03571	0.004461748
energy	1	244.22232	244.222321	976.88929	0.020361479
genre	3	5858.49471	1952.831571	7811.32628	0.008317154
danceability:energy	1	32.26765	32.267647	129.07059	0.055891869
danceability:genre	3	314.21270	104.737566	418.95026	0.035895334
energy:genre	3	80.51004	26.836680	107.34672	0.070803974
danceability:energy:genre	2	18.97115	9.485577	37.94231	0.114046058
Residuals	1	0.25000	0.250000	NA	NA

جدول فوق حاصل از انجام آزمون بیان شده است. با توجه به مقادیر p-value ها، نتایج بدین شرح تفسیر می‌شود:
Danceability، Energy و genre هر کدام به تنهایی تأثیر آماری قابل توجهی بر Popularity در سطح significance برابر ۰.۰۵ دارند. تاثیر متقابل بین Danceability:genre نیز در سطح significance برابر ۰.۰۵ قابل توجه است. تاثیر متقابل Danceability:Energy و Energy:genre با توجه به این‌که بین ۰.۰۵ و ۰.۱ قرار دارد، اندکی قابل توجه است. در حالی که اثر متقابل بین Danceability:Energy:genre قابل توجه

نیست و مقدار p-value مربوط به آن از ۰.۱ نیز بیشتر است.

۵-۱ زیربخش ۵

(برای حل این بخش، از کد R استفاده شده‌است.)

نتایج حاصل از اجرای رگرسیون خطی چندگانه‌ی خواسته شده، به شرح زیر است:

```
Residuals:
    Min      1Q  Median      3Q     Max
-45.604 -10.759   5.406  13.196  24.131

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -42.033     37.692  -1.115   0.2850
danceability   12.081      4.606   2.623   0.0211 *
energy          3.736      5.264   0.710   0.4905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.02 on 13 degrees of freedom
Multiple R-squared:  0.4583,    Adjusted R-squared:  0.3749
F-statistic: 5.499 on 2 and 13 DF,  p-value: 0.0186
```

نتیجه‌ی حاصل، نشان می‌دهد که بین Popularity (به عنوان متغیر وابسته) و Dance-ability (به عنوان یکی از متغیرهای مستقل) با مقدار t-value برابر ۲.۶۲۳ و p-value برابر ۰.۰۲۱۱ رابطه‌ی قابل توجه و معناداری با در نظر گرفتن سطح significance برابر ۰.۰۵ وجود دارد. این بدان معناست که به ازای هر افزایش یک واحدی در مقدار Danceability، در صورتی‌که بقیه‌ی متغیرها ثابت بمانند، انتظار می‌رود مقدار Popu-larity به اندازه‌ی ۱۲.۰۸۱ واحد افزایش یابد.

همچنین، مشاهده می‌شود که رابطه‌ی بین Popularity (به عنوان متغیر وابسته) و Energy (به عنوان یکی از متغیرهای مستقل) با مقدار t-value برابر ۰.۷۱ و p-value برابر ۰.۴۹۰۵ قابل توجه و معنادار نیست. این نشان می‌دهد که هنگامی که مقدار Danceability ثابت نگه داشته شود، Energy تأثیر آماری قابل توجه و معناداری بر Popularity ندارد.

significant بودن مدل به صورت کلی، با استفاده از مقدار آماره‌ی F تعیین می‌شود؛ که مقدار آن، این را نشان می‌دهد که آیا برازش کلی مدل، به طور قابل توجه و معناداری بهتر از حالت شانسی است یا خیر. حال، با توجه به این‌که مقدار آماره‌ی F برابر ۵.۴۹۹

است و مقدار p-value برابر ۰.۰۱۸۶ است، قابل توجه و معنادار بودن عملکرد کلی مدل، نشان داده شده‌است. همچنین، مقدار R-squared تعدیل شده برابر ۰.۳۷۴۹ است و این را نشان می‌دهد که مدل تولید شده، تقریباً ۳۷.۴۹ درصد از واریانس در مقدار Popularity را توضیح می‌دهد، که میزان متوسطی‌ست.
در کل، این نتایج نشان می‌دهد که برای پیش‌بینی Popularity موسیقی‌ها، Dance-ability نسبت به Energy عامل مهم‌تری‌ست. و مدل تولید شده، به‌طور کلی تا حدی دارای قدرت پیش‌بینی نسبی است؛ ولی احتمالاً عوامل دیگری نیز وجود دارد که بر Popularity اثر می‌گذارند.

۲ مقایسه‌ی مدل

۲-۱ زیربخش ۱

سه معیار بررسی شده برای بررسی عملکرد مدل‌ها در این مسئله، به شرح زیر تفسیر می‌شوند:
- معیار accuracy، نسبت آهنگ‌هایی که به‌طور دقیق توسط هر یک از سه مدل، به‌عنوان مشابه یا غیرمشابه به آهنگ‌های مورد علاقه‌ی کاربر، طبقه‌بندی شده‌اند، به همه‌ی آهنگ‌ها را نشان می‌دهد. مقدار بالای accuracy برای یک مدل، نشان دهنده‌ی این است، که مدل می‌تواند بخش زیادی از آهنگ‌ها را، به طور دقیق، به عنوان مشابه یا غیرمشابه به آهنگ‌های مورد علاقه‌ی کاربر طبقه‌بندی کند. که مقدار بالای آن، برای ایجاد لیست‌های پخش شخصی‌سازی‌شده و منعکس‌کننده‌ی ترجیحات موسیقی کاربر، مهم است.
- معیار precision، نسبت آهنگ‌های توصیه‌شده توسط سیستم به کاربر، که مشابه به آهنگ‌های مورد علاقه‌ی او هستند، به همه‌ی آهنگ‌های توصیه‌شده توسط سیستم به او را نشان می‌دهد. این معیار، توانایی سیستم برای توصیه‌ی دقیق آهنگ‌های مشابه به آهنگ‌های مورد علاقه‌ی کاربر، را اندازه گیری می‌کند. مقدار بالای آن، نشان دهنده‌ی این است، که مدل می‌تواند آهنگ‌هایی را توصیه کند که به احتمال زیاد مورد استقبال کاربر قرار می‌گیرند و کیفیت کلی لیست‌های پخش شخصی‌سازی‌شده را افزایش می‌دهد.
- معیار recall، نسبت آهنگ‌های توصیه‌شده توسط سیستم به کاربر، که مشابه

به آهنگ‌های مورد علاقه‌ی او هستند، به همه‌ی آهنگ‌هایی که در واقع مشابه به آهنگ‌های مورد علاقه‌ی کاربر هستند، از جمله آن‌هایی که توسط سیستم توصیه نشده‌اند، را نشان می‌دهد. این معیار، توانایی مدل برای شناسایی و توصیه‌ی درصد بالایی از آهنگ‌های مورد علاقه‌ی کاربر را اندازه گیری می‌کند. مقدار بالای آن، نشان دهنده‌ی آن است، که مدل می‌تواند به طور موثر بخش بزرگی از آهنگ‌های مشابه به آهنگ‌های مورد علاقه‌ی کاربر را، شناسایی و توصیه کند؛ که برای اطمینان از این‌که آیا لیست‌های پخش شخصی‌سازی‌شده واقعاً منعکس‌کننده‌ی سلیقه‌ی موسیقی کاربر هستند، مهم است.

با توجه به این‌که بیان شده که هدف مدل این است که با گنجاندن آهنگ‌هایی که بسیار شبیه به آهنگ‌های مورد علاقه‌ی کاربران هستند، فهرست‌های پخش متناسب با تک تک آن‌ها ایجاد شود، و لیست‌های پخش باید دقیقاً مشابه آنچه کاربران می‌خواهند بشنوند، باشد، و با توجه به مفاهیمی که بالا برای ۳ معیار بیان شد، معیار precision به عنوان معیار مورد بررسی برای ارزیابی مدل‌ها، در نظر گرفته می‌شود. و آزمون مورد نظر، بر روی این معیار انجام می‌شود.

از آزمون Kruskal Wallis به عنوان آزمون non-parametric (غیرپارامتری) مورد نظر استفاده خواهیم‌کرد. این آزمون یک جایگزین غیرپارامتری برای آزمون ANOVA یک طرفه است. غیرپارامتری بودن یک آزمون بدین معناست که فرضی درباره‌ی توزیع داده‌ها در نظر نمی‌گیرد. گاهاً به آن one-way ANOVA on ranks نیز گفته می‌شود؛ زیرا به جای استفاده از خود داده‌ها، از rank آن‌ها استفاده می‌کند. در این آزمون، متفاوت بودن میانه‌های گروه‌ها بررسی می‌شود. در واقع، این را بیان می‌کند که آیا تفاوت معنی‌دار و قابل توجهی بین گروه‌ها وجود دارد یا خیر. و این را بیان نمی‌کند که کدام گروه‌ها متفاوت هستند. برای آن، باید از یک آزمون دیگر مثل Post-Hoc استفاده کرد.
این آزمون، انتخاب مناسبی برای این حل این مسئله است؛ زیرا توزیع داده‌ها بیان نشده‌است و تنها ۵ نمونه در هر مدل وجود دارد، که تأیید فرض نرمال بودن داده‌ها (یا پیروی داده‌ها از هر توزیع دیگری) را دشوار می‌کند.

فرض صفر و فرض جایگزین:
-فرض صفر، این را بیان می‌کند که میانه‌های precision های هر سه مدل، با هم برابر است.

$$H_0 : \quad median_A = median_B = median_C \tag{۱۹}$$

-فرض جایگزین، این را بیان می‌کند که میانه‌های precision های حداقل یکی از مدل‌ها، با بقیه برابر نیست.

$$\begin{aligned} H_A : \quad & median_A \neq median_B \\ or \quad & median_A \neq median_C \\ or \quad & median_B \neq median_C \end{aligned} \tag{۲۰}$$

- گام اول: مقادیر precision های مربوط به تمامی مدل‌ها، در یک مجموعه، با هم ترکیب و به صورت صعودی مرتب می‌شوند. و به آن‌ها rank متناظرشان، نسبت داده می‌شود.

rank	precision
۱	۰.۸۳
۲	۰.۸۴
۳	۰.۸۵
۴ -> ۴.۵	۰.۸۶
۵ -> ۴.۵	۰.۸۶
۶	۰.۸۷
۷	۰.۸۹
۸	۰.۹۰
۹	۰.۹۲
۱۰ -> ۱۰.۵	۰.۹۳
۱۱ -> ۱۰.۵	۰.۹۳
۱۲	۰.۹۴
۱۳	۰.۹۵
۱۴	۰.۹۶
۱۵	۰.۹۷

- گام دوم: rank های نظیر مقادیر precision های مربوط به هر یک از مدل‌ها، جمع می‌شود.

$$Model_A : 0.95, 0.93, 0.96, 0.94, 0.97 : 13 + 10.5 + 14 + 12 + 15 = 64.5 \quad (۲۱)$$

$$Model_B : 0.85, 0.86, 0.84, 0.83, 0.86 : 3 + 4.5 + 2 + 1 + 4.5 = 15 \quad (۲۲)$$

$$Model_C : 0.90, 0.87, 0.92, 0.89, 0.93 : 8 + 6 + 9 + 7 + 10.5 = 40.5 \quad (۲۳)$$

- گام سوم: مقدار آماره‌ی H با استفاده از فرمول زیر محاسبه می‌شود:
(پارامتر n نشان‌دهنده‌ی مجموع سایز تمام نمونه‌هاست. c نشان دهنده‌ی تعداد نمونه‌هاست. Tj مجموع rank های نمونه‌ی j ام است. nj سایز نمونه‌ی j ام است.)

$$\begin{aligned} H &= \left[\frac{12}{n(n+1)}\sum_{i=1}^{c}\frac{T_j^2}{n_j}\right] - 3(n+1) = \frac{12}{15(16)}\left(\frac{64.5^2}{5} + \frac{15^2}{5} + \frac{40.5^2}{5}\right) - 3(16) \\ &= \frac{12}{240}\left(\frac{4160.25}{5} + \frac{225}{5} + \frac{1640.25}{5}\right) - 48 = 0.05(832.05 + 45 + 328.05) - 48 \\ &= 0.05(1205.1) - 48 = 60.255 - 48 = 12.255 \end{aligned} \quad (۲۴)$$

- گام چهارم: مقدار بحرانی chi-square با درجه آزادی c-۱ (۳-۱=۲) و سطح signifi-cance برابر ۰.۰۵ برابر ۵.۹۹۱۵ است.

- گام پنجم: مقدار بحرانی chi-square با مقدار بدست آمده برای H مقایسه می‌شود. با توجه به این‌که مقدار بحرانی chi-square کمتر از آماره‌ی H است، فرض صفر رد می‌شود. پس نمی‌توان ادعا کرد که میانه‌های precision های هر سه مدل، با هم برابر است.

در نتیجه، با توجه به آزمون انجام شده، حداقل یکی از مدل‌های مدنظر، دارای عملکرد به طور قابل توجه متفاوتی نسبت به بقیه‌ی مدل‌هاست. اما این آزمون به ما نمی‌گوید که کدام یک، متفاوت است یا بهتر است. حالا که فهمیدیم عملکرد آن‌ها متفاوت است، از آزمون post-hoc برای فهمیدن این‌که کدام متفاوت است، استفاده می‌کنیم:

- گام اول: میانگین و انحراف معیار precision های هر یک از مدل‌ها محاسبه می‌شود:

$$mean_A = \frac{0.95 + 0.93 + 0.96 + 0.94 + 0.97}{5} = \frac{4.75}{5} = 0.95 \quad , \quad sd_A = 0.0158 \tag{۲۵}$$

$$mean_B = \frac{0.85 + 0.86 + 0.84 + 0.83 + 0.86}{5} = \frac{4.24}{5} = 0.848 \quad , \quad sd_B = 0.0130 \tag{۲۶}$$

$$mean_C = \frac{0.90 + 0.87 + 0.92 + 0.89 + 0.93}{5} = \frac{4.51}{5} = 0.902 \quad , \quad sd_C = 0.0238 \tag{۲۷}$$

- گام دوم: میانگین کلی precision های تمام مدل‌ها محاسبه می‌شود:

$$mean = \frac{0.95 + 0.848 + 0.902}{3} = 0.9 \tag{۲۸}$$

- گام سوم: مقدار SSG محاسبه می‌شود:

$$SSG = 5 * ((0.95 - 0.9)^2 + (0.848 - 0.9)^2 + (0.902 - 0.9)^2) = 0.026 \tag{۲۹}$$

- گام چهارم: درجه آزادی مربوط به SSG محاسبه می‌شود:

$$df(SSG) = NumberOfGroups - 1 = 3 - 1 = 2 \tag{۳۰}$$

- گام پنجم: مقدار SST محاسبه می‌شود:

$$\begin{aligned} SST = &(0.95 - 0.9)^2 + (0.93 - 0.9)^2 + (0.96 - 0.9)^2 + (0.94 - 0.9)^2 + (0.97 - 0.9)^2 \\ &+(0.85 - 0.9)^2 + (0.86 - 0.9)^2 + (0.84 - 0.9)^2 + (0.83 - 0.9)^2 + (0.86 - 0.9)^2 \\ &+(0.90 - 0.9)^2 + (0.87 - 0.9)^2 + (0.92 - 0.9)^2 + (0.89 - 0.9)^2 + (0.93 - 0.9)^2 \\ &= 0.03 \end{aligned} \tag{۳۱}$$

- گام ششم: مقدار SSE محاسبه می‌شود:

$$SSE = 0.03 - 0.026 = 0.004 \quad (۳۲)$$

- گام هفتم: درجه آزادی Error محاسبه می‌شود:

$$df(Error) = TotalNumberOfObservations - NumberOfGroups = 15 - 3 = 12 \quad (۳۳)$$

- گام هشتم: مقدار MSG محاسبه می‌شود:

$$MSG = SSG/df(SSG) = 0.026/2 = 0.013 \quad (۳۴)$$

- گام نهم: مقدار MSE محاسبه می‌شود:

$$MSE = SSE/df(Error) = 0.004/12 = 0.0003 \quad (۳۵)$$

- گام دهم: مقدار بحرانی روش Tukey's HSD برای سطح significance برابر ۰.۰۵ و تعداد گروه‌ها برابر ۳ و درجه آزادی SSW برابر ۱۲ با توجه به جدول مربوطه، برابر ۳.۷۷ است.

$$q = 3.77 \quad (۳۶)$$

- گام یازدهم: مقدار HSD با استفاده از فرمول زیر محاسبه می‌شود:

$$HSD = q * \sqrt{\frac{MSE}{n}} = 3.77 * \sqrt{\frac{0.0003}{15}} = 3.77 * \sqrt{0.00002} = 3.77 * 0.0044 = 0.0165 \quad (۳۷)$$

- گام دوازدهم: با توجه به این‌که می‌دانیم تفاوت قابل توجهی میان عملکرد مدل‌ها وجود دارد، اختلاف میان میانگین‌های هر کدام از زوج مدل‌های ممکن محاسبه می‌شود:

$$|mean_A - mean_B| = |0.95 - 0.848| = 0.102 \quad (۳۸)$$

$$|mean_A - mean_C| = |0.95 - 0.902| = 0.048 \quad (۳۹)$$

$$|mean_B - mean_C| = |0.848 - 0.902| = 0.054 \quad (۴۰)$$

- گام سیزدهم: مقدار standard error مربوط به اختلاف میانگین‌های زوج مدل‌ها محاسبه می‌شود. و با تقسیم اختلاف میانگین مطلق هر کدام از زوج مدل‌ها بر این standard error مقدار آماره‌ی T آن‌ها محاسبه می‌شود:

$$StandardError = \sqrt{2 * \frac{MSE}{n}} = \sqrt{2 * \frac{0.0003}{15}} = \sqrt{0.00004} = 0.0063 \quad (۴۱)$$

$$T_{(A,B)} = \frac{0.102}{0.0063} = 16.19 \quad (۴۲)$$

$$T_{(A,C)} = \frac{0.048}{0.0063} = 7.619 \quad (۴۳)$$

$$T_{(B,C)} = \frac{0.054}{0.0063} = 8.571 \quad (۴۴)$$

- گام چهاردهم: برای بررسی قابل توجه بودن تفاوت هر زوج مدل، مقدار آماره‌ی T آن‌ها، با مقدار HSD مقایسه می‌شود:
مقدار آماره‌ی T مربوط به همه‌ی زوج مدل‌ها، از مقدار HSD محاسبه شده بیشتر

است. در نتیجه، عملکرد همه‌ی مدل‌ها به طور قابل توجهی با هم متفاوت است.

* با توجه به محاساباتی که انجام شد، و با توجه به مقادیر precision ها، عملکرد مدل A از بقیه‌ی مدل‌ها بهتر است.

۲-۲ زیربخش ۲

با توجه به این‌که متناسب با هدف سیستم، معیار precision به عنوان معیار بررسی عملکرد مدل‌ها انتخاب شد، و مدل A به عنوان مدل با بهترین عملکرد تشخیص داده‌شد، باید precision های داده شده در این سوال را به عنوان precision های جدید مدل A پس از ایجاد تنظیمات جدید در نظر گرفت. و با precision های قبلی مدل A مقایسه کرد.
به طور واضح، این precision ها نسبت به قبل کم‌تر شده‌اند. و مدل A در حالت قبلی، دارای عملکرد بهتری بود. چرا که، اگر به صورت نظیر به نظیر، مقادیر جدید precision را با مقادیر قبلی آن مقایسه کنیم، در تمام ۵ بار، کاهش مشاهده می‌شود.

۳ مقایسه‌ی سن طرفداران

۳-۱ زیربخش ۱

برای بررسی نرمال بودن/نبودن توزیع، از آزمون Shapiro-Wilk استفاده می‌شود. در این آزمون، آماره با استفاده از فرمول زیر محاسبه می‌شود. که در آن $x_{(i)}$ ، i امین آماره‌ی ترتیبی است. $\overline{x}$ میانگین نمونه است و ثابت‌های a_i با استفاده از فرمول پایین‌تر به دست می‌آیند.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad (۴۵)$$

$$(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{(m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}} \quad (۴۶)$$

$$m = (m_1, \ldots, m_n)^{\mathsf{T}}$$

فرض صفر و فرض جایگزین:

- توزیع نرمال است.
- توزیع نرمال نیست.

در صورتی که این آماره از سطح significance برابر ۰.۰۵ کم‌تر باشد، می‌توان فرض صفر یا همان نرمال بودن را رد کرد.

نتیجه‌ی حاصل از اجرای این آزمون به شرح زیر است:

```
        Shapiro-Wilk normality test

data:  men_age
W = 0.94604, p-value = 0.02351

        Shapiro-Wilk normality test

data:  women_age
W = 0.95655, p-value = 0.06375
```

با توجه به مقادیر p-value هر یک از دو گروه زنان و مردان، توزیع سن طرفداران مرد، از توزیع نرمال پیروی نمی‌کند و توزیع سن طرفداران زن، از توزیع نرمال پیروی می‌کند.

۳-۲ زیربخش ۲

برای این‌که امکان استفاده از آزمون‌های پارامتری را داشته باشیم، باید یک سری شرایط در مسئله برقرار باشد.
برای استفاده از آزمون‌های پارامتری، باید این شرایط برقرار باشند:

۱. توزیع داده‌ها نرمال باشد.
باتوجه به پاسخ بخش قبل، چون توزیع سن طرفداران مرد، نرمال نبود، امکان استفاده از آزمون های پارامتری وجود ندارد.

۲. گروه‌های مختلف دارای واریانس برابر باشند.
برای بررسی این شرط، از آزمون Bartlett استفاده می‌شود.
فرض صفر و فرض جایگزین:
- واریانس گروه‌ها برابر است.
- واریانس گروه‌ها برابر نیست.
اگر مقدار p-value به دست آمده از آزمون کم‌تر از سطح significance باربر ۰.۰۵ باشد، فرض صفر رد می‌شود و این نتیجه حاصل می‌شود که واریانس‌های گروه‌ها به طور قابل توجهی متفاوت است.
نتیجه‌ی حاصل از اجرای این آزمون به شرح زیر است:

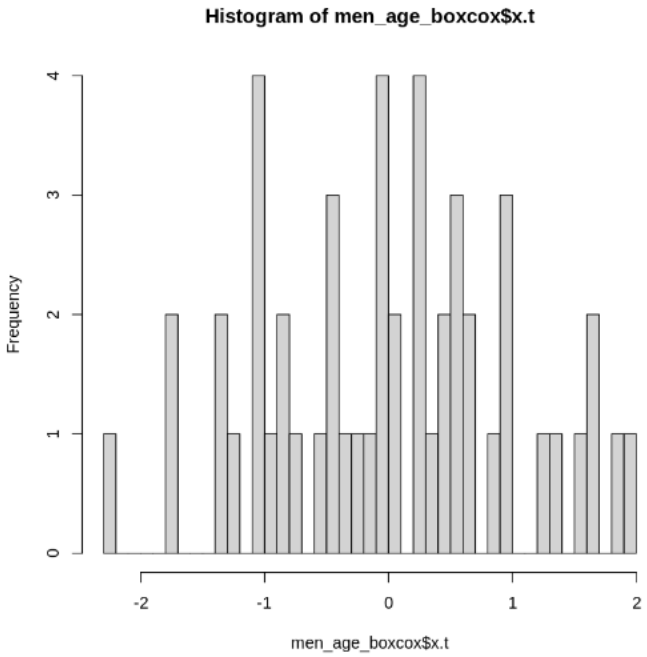
```
        Bartlett test of homogeneity of variances

data:  list(men_age, women_age)
Bartlett's K-squared = 1.7956, df = 1, p-value = 0.1802
```

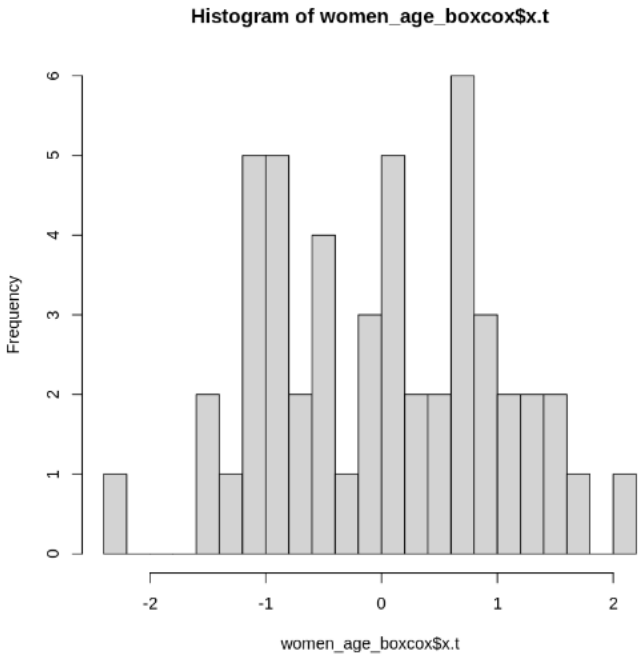
با توجه به مقدار p-value نمی‌توان فرض صفر را رد کرد. و احتمالاً واریانس دو گروه مردان و زنان برابر نیست.

۳. داده‌ها مستقل از هم باشند.
سن طرفداران، یک متغیر مستقل است.

۳-۳ زیربخش ۳

برای نرمال کردن داده‌ها، از boxcox استفاده شد.
مقدار p-value گروه‌ها و نمودار هیستوگرام مربوط به آن‌ها، پس از نرمال سازی، به شرح زیر است:





```
        Shapiro-Wilk normality test

data:  men_age_boxcox$x.t
W = 0.98669, p-value = 0.8411

        Shapiro-Wilk normality test

data:  women_age_boxcox$x.t
W = 0.98056, p-value = 0.576
```

همانطور که ملاحظه می‌شود، داده‌ها نرمال شده اند.
حال شرایط انجام تست پارامتری برقرار است.

۳-۴ زیربخش ۴

از آزمون T استفاده شد.
فرض صفر و فرض جایگزین:
- تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود ندارد.
- تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود دارد.
نتایج آزمون به شرح زیر است:

```
        Welch Two Sample t-test

data:  men_age_boxcox$x.t and women_age_boxcox$x.t
t = -3.0594e-15, df = 98, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3968935  0.3968935
sample estimates:
    mean of x     mean of y
-2.919995e-16  3.198830e-16
```

همانطور که ملاحظه می‌شود، مقدار p-value برابر ۱ بدست آمد، که از سطح sig-nificance برابر ۰.۰۵ بیشتر است؛ در نتیجه نمی‌توان فرض صفر را رد کرد. و احتمالاً تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود ندارد.

۳-۵ زیربخش ۵

به عنوان آزمون ناپارامتری از آزمون Mann-Whitney U استفاده شد.
فرض صفر و فرض جایگزین:

- تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود ندارد.
- تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود دارد.

نتایج آزمون به شرح زیر است:

```
        Wilcoxon rank sum test with continuity correction

data:  men_age and women_age
W = 1224, p-value = 0.8604
alternative hypothesis: true location shift is not equal to 0
```

همانطور که ملاحظه می‌شود، مقدار p-value برابر ۰.۸ بدست آمد، که از سطح sig-nificance برابر ۰.۰۵ بیشتر است؛ در نتیجه نمی‌توان فرض صفر را رد کرد. و احتمالاً تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود ندارد.

مقایسه‌ی نتایج دو آزمون:

هر دو تست پارامتری و ناپارامتری به نتیجه‌های مشابهی منجر می‌شوند که نشان می‌دهد تفاوت قابل توجهی میان سن دو گروه طرفداران زن و مرد وجود ندارد. مقادیر p-value برای هر دو آزمون، بالا است؛ که نشان می‌دهد شواهد آماری کافی برای رد کردن فرض صفر وجود ندارد.

مقایسه‌ی قدرت دو تست:

از نظر قدرت آماری، تست های پارامتری مانند آزمون T معمولاً قدرت بیشتری نسبت به تست های ناپارامتری مانند آزمون Mann-Whitney U دارند. با این حال، در این مورد، هر دو آزمون به نتیجه‌های مشابهی منجر شدند.

۶-۳ زیربخش ۶

با توجه به اندازه‌ی نمونه‌ها (۵۰ طرفدار مرد و ۵۰ طرفدار زن)، و اینکه فرض نرمال بودن داده‌ها برآورده نمی‌شود، استفاده از آزمون ناپارامتری Mann-Whitney U در این موقعیت مناسب‌تر است. این آزمون نیازی به فرض نرمال بودن داده‌ها ندارد و در مقابل نقض سایر فرضیات نیز مقاوم است. علاوه بر این، نتیجه‌ی به دست آمده از این آزمون، با نتیجه‌ی آزمون T تطابق دارد.

۴ همکاری هنرمندان در Spotify

(فایل حاوی کد R ضمیمه شده‌است.)

۴-۱ زیربخش ۱

دو فایل مربوط به گره‌ها و یال‌های گراف با استفاده از کد R خوانده شد و گراف نظیر آن شکل داده‌شد:

```
IGRAPH 79e3d12 UN-- 12905 321566 --
+ attr: name (v/c), isdone (v/c), spotifyid (v/c), genres (v/c),
| popularity (v/n), followers (v/n), histogram (v/c), num_release
| (v/n), first_release (v/c), last_release (v/c), network_rank (v/n),
| songid (e/n), song (e/c)
+ edges from 79e3d12 (vertex names):
 [1] Lil Wayne--Drake            Drake    --Nicki Minaj
 [3] Rihanna  --Drake            Rihanna  --Drake
 [5] Lil Wayne--Drake            Drake    --Nicki Minaj
 [7] Rihanna  --Drake            Rihanna  --Drake
 [9] Rick Ross--Drake            Drake    --Waka Flocka Flame
+ ... omitted several edges
```

۴-۲ زیربخش ۲

PageRank یک الگوریتم رتبه بندی است که به هر گره یک امتیاز بر اساس تعداد، کیفیت و اهمیت سایر گره‌ها که به آن لینک می‌دهند، اختصاص می‌دهد؛ با این فرض که گره‌های مهم‌تر احتمالاً لینک‌های بیشتری را از گره‌های دیگر دریافت می‌کنند. در زمینه‌ی گراف همکاری هنرمندان در Spotify می‌توانیم از الگوریتم PageRank برای رتبه‌بندی هنرمندان بر اساس تعداد و کیفیت همکاری‌هایی که با هنرمندان دیگر دارند استفاده کنیم. با این تفسیر که اگر یک هنرمند، با هنرمندان مهم دیگر همکاری کند، اهمیت بیشتری دارد.

مقدار PageRank برای گره‌ها در گراف محاسبه شد. Wolfgang Amadeus Mozart با PageRank برابر با ۰.۰۰۳۸۳۹ و Ludwig van Beethoven با PageRank برابر با ۰.۰۰۳۶۹۵ دو هنرمند با بیشترین PageRank در این گراف هستند. همچنین، دو هنرمند با نام‌های D.J. Mixxy B و Adrien Lamont از جمله هنرمندان با کم‌ترین PageRank هستند.

مرکزیت درجه، برابر است با تعداد لینک‌هایی که آن گره دریافت می‌کند (در واقع، درجه‌ی آن گره). در نتیجه، هر چه درجه‌ی یک گره بالاتر باشد، آن گره مرکزی‌تر است. در زمینه‌ی گراف همکاری هنرمندان در Spotify اگر یک هنرمند با هنرمندان بیشتری همکاری کند، مرکزیت بیشتری دارد.
مقدار مرکزیت درجه برای گره‌ها در گراف محاسبه شد. Ludwig van Beethoven با مرکزیت درجه برابر با ۱۲۴۱۲ و Wolfgang Amadeus Mozart با مرکزیت درجه برابر با ۱۲۳۶۴ دو هنرمند با بیشتربن مرکزیت درجه در این گراف هستند. همچنین، دو هنرمند با نام‌های D.J. Mixxy B و Adrien Lamont از جمله هنرمندان با کم‌ترین مرکزیت درجه (۰) هستند.

۳-۴ زیربخش ۳

ده هنرمند با بیشترین PageRank به شرح زیر هستند:

	name
	<chr>
1346	Wolfgang Amadeus Mozart
271	Ludwig van Beethoven
559	Johann Sebastian Bach
1439	Johannes Brahms
1446	London Symphony Orchestra
3899	Пётр Ильич Чайковский
1875	Antonio Vivaldi
1458	Franz Schubert
62	Snoop Dogg
8301	Farruko

این هنرمندان در واقع دارای بیشتربن تعداد همکاری با دیگر هنرمندان هستند و همچنین این همکاری‌هایی که دارند، با هنرمندانی‌ست که آن‌ها نیز با هنرمندان زیادی همکاری دارند.

این هنرمندان با PageRank بالا، احتمالاً هنرمندانی هستند که دارای محبوبیت بالایی در Spotify هستند، تعداد دنبال‌کنندگان زیادی در Spotify دارند، و همچنین تعداد آهنگ‌های زیادی منتشر کرده‌اند. این ویژگی‌ها، با استفاده از ۳ attribute با نام‌های popularity، followers و num release در مجموعه داده‌ها ظاهر شده‌اند. این ۳ attribute احتمالاً دارای بیشترین تاثیر روی PageRank هستند.

برای اثبات فرض بیان شده، یک رگرسیون خطی چندگانه بر روی ۳ متغیر بیان شده (به عنوان متغیرهای مستقل) و PageRank (به عنوان متغیر وابسته) اجرا شد. نتایج آن به شرح زیر است:

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.602e-03 -3.518e-05 -1.110e-05  2.211e-05  2.469e-03

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.204e-05  2.182e-06   10.10   <2e-16 ***
popularity  6.418e-07  4.699e-08   13.66   <2e-16 ***
followers   3.566e-12  3.429e-13   10.40   <2e-16 ***
num_release 1.561e-06  2.562e-08   60.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.944e-05 on 12901 degrees of freedom
Multiple R-squared:  0.298,     Adjusted R-squared:  0.2978
F-statistic:  1825 on 3 and 12901 DF,  p-value: < 2.2e-16
```

نتیجه‌ی حاصل نشان می‌دهد که بین PageRank (به عنوان متغیر وابسته) و popularity (به عنوان یکی از متغیرهای مستقل) با مقدار t-value برابر ۱۳.۶۶ و p-value بسیار کوچک رابطه‌ی قابل توجه و معناداری با در نظر گرفتن سطح significance برابر ۰.۰۵ وجود دارد. این بدان معناست که به ازای هر افزایش یک واحدی در مقدار popularity در صورتی‌که بقیه‌ی متغیرها ثابت بمانند، انتظار می‌رود مقدار PageRank به اندازه‌ی ۰.۰۰۰۰۰۰۶۴۱۸ واحد افزایش یابد.
همچنین، مشاهده می‌شود که رابطه‌ی بین PageRank (به عنوان متغیر وابسته) و followers (به عنوان یکی از متغیرهای مستقل) با مقدار t-value برابر ۱۰.۴ و p-value بسیار کوچک، قابل توجه و معنادار است. این نشان می‌دهد که هنگامی که بقیه‌ی متغیرها ثابت نگه داشته شوند، followers تأثیر آماری قابل توجه و معناداری بر PageRank دارد. این بدان معناست که به ازای هر افزایش یک واحدی در مقدار

followers، در صورتی‌که بقیه‌ی متغیرها ثابت بمانند، انتظار می‌رود مقدار PageRank به اندازه‌ی ۰.۰۰۰۰۰۰۰۰۰۰۰۳۵۶۶ واحد افزایش یابد.
همینطور، رابطه‌ی بین PageRank (به عنوان متغیر وابسته) و num release (به عنوان یکی از متغیرهای مستقل) با مقدار t-value برابر ۶۰.۹۱ و p-value بسیار کوچک، قابل توجه و معنادار است. این نشان می‌دهد که هنگامی که بقیه‌ی متغیرها ثابت نگه داشته شود، num release تأثیر آماری قابل توجه و معناداری بر PageRank دارد. این بدان معناست که به ازای هر افزایش یک واحدی در مقدار num release در صورتی‌که بقیه‌ی متغیرها ثابت بمانند، انتظار می‌رود مقدار PageRank به اندازه‌ی ۰.۰۰۰۰۰۱۵۶۱ واحد افزایش یابد.
significant بودن مدل به صورت کلی، با استفاده از مقدار آماره‌ی F تعیین می‌شود؛ که مقدار آن، این را نشان می‌دهد که آیا برازش کلی مدل، به طور قابل توجه و معناداری بهتر از حالت شانسی است یا خیر. حال، با توجه به این‌که مقدار آماره‌ی F برابر ۱۸۲۵ است و مقدار p-value بسیار کوچک است، قابل توجه و معنادار بودن عملکرد کلی مدل، نشان داده شده‌است. همچنین، مقدار R-squared تعدیل شده برابر ۰.۲۹۷۸ است و این را نشان می‌دهد که مدل تولید شده، تقریباً ۲۹.۷۸ درصد از واریانس در مقدار PageRank را توضیح می‌دهد، که میزان متوسط رو به پایینی‌ست.
در کل، این نتایج نشان می‌دهد که فرض اولیه درست بوده است و این متغیرها تاثیر قابل توجهی بر مقدار PageRank دارند. و از بین آن‌ها، متغیر num release برای پیش‌بینی PageRank عامل مهم‌تری‌ست. و مدل تولید شده، به‌طور کلی تا حدی دارای قدرت پیش‌بینی نسبی است؛ ولی احتمالاً عوامل دیگری نیز وجود دارد که بر PageRank اثر می‌گذارند.

همچنین، برای بررسی رابطه و همبستگی هر یک از این متغیرها با متغیر PageRank، مقادیر correlation coefficient مربوطه محاسبه شد:

```
Correlation between PageRank and popularity:  0.2711734
Correlation between PageRank and followers:  0.2364538
Correlation between PageRank and num_release:  0.5243757
```

بیشترین همبستگی بین PageRank و num release مشاهده می‌شود، که نشان می‌دهد این متغیر احتمالاً مهم‌ترین پیش‌بینی‌کننده برای PageRank در بین این سه متغیر مستقلِ مورد نظر باشد.

(فرض صفر و فرض جایگزین در این دو تست آماری:
- فرض صفر: هیچ رابطه‌ی قابل توجه و معناداری میان این متغیرهای مستقل بیان شده و PageRank وجود ندارد.
- فرض جایگزین: میان این متغیرهای مستقل بیان شده و PageRank رابطه‌ی قابل توجه و معناداری وجود دارد.
با توجه به نتایج، فرض صفر رد شد. و احتمالا میان این متغیرهای مستقل بیان شده و PageRank رابطه‌ی قابل توجه و معناداری وجود دارد.)

به طور کلی، نتایج حاصل از این دو تست آماری انجام شده را می‌توان بدین شرح تفسیر کرد:
برای دو متغیر popularity و followers با توجه به این‌که رگرسیون چندگانه وجود روابط قابل توجه و معناداری را نشان می‌دهد، اما مقدار correlation مربوط به آن‌ها چندان خوب نیست، این مسئله برداشت می‌شود که بین این دو متغیر مستقل و PageRank به عنوان متغیر وابسته، از نظر آماری روابط معنادار و قابل توجهی وجود دارد؛ اما این روابط، چندان قوی نیستند. و برای متغیر num release با توجه به این‌که هم رگرسیون چندگانه، وجود رابطه‌ی قابل توجه و معناداری را نشان می‌دهد، و هم مقدار correlation مربوط به آن نسبتاً خوب است، این مسئله برداشت می‌شود که بین این دو متغیر مستقل و PageRank به عنوان متغیر وابسته، از نظر آماری رابطه‌ی معنادار و قابل توجهی وجود دارد؛ و این رابطه، نسبتاً قوی است.
همچنین، correlation لزوماً به معنای علّیت نیست؛ بنابراین حتی اگر correlation ها خیلی قوی نباشند، به این معنا نیست که رابطه‌ی علّی بین متغیرهای مستقل و وابسته وجود ندارد.

۴-۴ زیربخش ۴

تعداد همکاری‌های هنرمندان، در واقع همان درجه‌ی گره‌ی مربوط به آن‌ها در گراف است؛ که مقدار آن با مرکزیت درجه که قبلاً محاسبه شد، برابر است. در نتیجه، هدف این بخش، بررسی این مسئله است که آیا ارتباط قابل توجهی میان متغیر degree (تعداد همکاری‌های هنرمند) و متغیر popularity (محبوبیت هنرمند در Spotify)

وجود دارد.
برای بررسی قابل توجه بودن/نبودن ارتباط میان دو متغیر بیان شده، از مدل رگرسیون خطی و محاسبه‌ی correlation آن‌ها استفاده شد.
نتیجه‌ی مربوط به مدل رگرسیون خطی:

```
Residuals:
    Min      1Q  Median      3Q     Max
 -103.6   -55.1   -39.7   -12.3 12326.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3479     6.3199  -0.213    0.831
popularity    1.1406     0.1282   8.895   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 296.9 on 12903 degrees of freedom
Multiple R-squared:  0.006095,  Adjusted R-squared:  0.006018
F-statistic: 79.12 on 1 and 12903 DF,  p-value: < 2.2e-16
```

تیجه‌ی حاصل نشان می‌دهد که بین degree به عنوان متغیر وابسته و popularity به عنوان متغیر مستقل، با مقدار t-value برابر ۸.۸۹۵ و p-value بسیار کوچک رابطه‌ی قابل توجه و معناداری با در نظر گرفتن سطح significance برابر ۰.۰۵ وجود دارد. این بدان معناست که به ازای هر افزایش یک واحدی در مقدار popularity انتظار می‌رود مقدار degree به اندازه‌ی ۱.۱۴۰۶ واحد افزایش یابد.
significant بودن مدل به صورت کلی، با استفاده از مقدار آماره‌ی F تعیین می‌شود؛ که مقدار آن، این را نشان می‌دهد که آیا برازش کلی مدل، به طور قابل توجه و معناداری بهتر از حالت شانسی است یا خیر. حال، با توجه به اینکه مقدار آماره‌ی F برابر ۷۹.۱۲ است و مقدار p-value بسیار کوچک است، قابل توجه و معنادار بودن عملکرد کلی مدل، نشان داده شده‌است. همچنین، مقدار R-squared تعدیل شده برابر ۰.۰۰۶ است و این را نشان می‌دهد که مدل تولید شده، تقریباً ۰.۶ درصد از واریانس در مقدار degree را توضیح می‌دهد، که میزان کمی‌ست.

نتیجه‌ی مربوط به محاسبه‌ی correlation بین این دو متغیر:

```
Correlation between degree and popularity:  0.07806795
```

مقدار بسیار کم محاسبه شده، همبستگی بسیار کم میان این دو متغیر را نشان می‌دهد.

به طور کلی، نتایج حاصل از این دو تست آماری انجام شده را می‌توان بدین شرح تفسیر کرد:
با توجه به این‌که رگرسیون خطی وجود روابط قابل توجه و معناداری را نشان می‌دهد، اما مقدار correlation مربوط به آن‌ها چندان خوب نیست، این مسئله برداشت می‌شود که بین این دو متغیر، از نظر آماری رابطه‌ی معنادار و قابل توجهی وجود دارد؛ اما این رابطه، چندان قوی نیست.
همچنین، correlation لزوماً به معنای علّیت نیست؛ بنابراین حتی اگر correlation خیلی قوی نباشد، به این معنا نیست که رابطه‌ی علّی بین متغیرهای مستقل و وابسته وجود ندارد.

۵-۴ زیربخش ۵

- ارتباط میان تجربه‌ی کاری و PageRank:
برای محاسبه‌ی تجربه‌ی کاری، بازه‌ی زمانی میان اولین انتشار تا اخرین انتشار هر شخص را به عنوان تجربه‌ی کاری او در نظر می‌گیریم.
برای بررسی قابل توجه بودن/نبودن ارتباط میان دو متغیر بیان شده، از مدل رگرسیون خطی و محاسبه‌ی correlation آن‌ها استفاده شد.
نتیجه‌ی مربوط به مدل رگرسیون خطی:

```
Residuals:
       Min         1Q     Median         3Q        Max
-0.0002620 -0.0000402 -0.0000196  0.0000172  0.0036568

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.315e-05  1.280e-06   41.52   <2e-16 ***
experience  5.000e-09  1.621e-10   30.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001145 on 12903 degrees of freedom
Multiple R-squared:  0.06866,   Adjusted R-squared:  0.06859
F-statistic: 951.3 on 1 and 12903 DF,  p-value: < 2.2e-16
```

تیجه‌ی حاصل نشان می‌دهد که بین PageRank به عنوان متغیر وابسته و تجربه کاری به عنوان متغیر مستقل، با مقدار t-value برابر ۳۰.۸۴ و p-value بسیار کوچک رابطه‌ی قابل توجه و معناداری با در نظر گرفتن سطح significance برابر ۰.۰۵ وجود دارد. این بدان معناست که به ازای هر افزایش یک واحدی در مقدار تجربه کاری انتظار می‌رود مقدار PageRank به اندازه‌ی ۰.۰۰۰۰۰۰۰۰۵ واحد افزایش یابد.
significant بودن مدل به صورت کلی، با استفاده از مقدار آماره‌ی F تعیین می‌شود؛ که مقدار آن، این را نشان می‌دهد که آیا برازش کلی مدل، به طور قابل توجه و معناداری بهتر از حالت شانسی است یا خیر. حال، با توجه به اینکه مقدار آماره‌ی F برابر ۹۵۱.۳ است و مقدار p-value بسیار کوچک است، قابل توجه و معنادار بودن عملکرد کلی مدل، نشان داده شده‌است. همچنین، مقدار R-squared تعدیل شده برابر ۰.۰۶۸ است و این را نشان می‌دهد که مدل تولید شده، تقریباً ۶.۸ درصد از واریانس در مقدار PageRank را توضیح می‌دهد، که میزان کمی‌ست.

نتیجه‌ی مربوط به محاسبه‌ی correlation بین این دو متغیر:

```
Correlation between pagerank and work experience:  0.2620375
```

مقدار متوسط رو به پایین محاسبه شده، وجود مقدار همبستگی به میزان متوسط میان این دو متغیر را نشان می‌دهد.

به طور کلی، نتایج حاصل از این دو تست آماری انجام شده را می‌توان بدین شرح تفسیر کرد:
با توجه به این‌که رگرسیون خطی وجود روابط قابل توجه و معناداری را نشان می‌دهد، اما مقدار correlation مربوط به آن‌ها چندان خوب نیست، این مسئله برداشت می‌شود که بین این دو متغیر، از نظر آماری رابطه‌ی معنادار و قابل توجهی وجود دارد؛ اما این رابطه، چندان قوی نیست.
همچنین، correlation لزوماً به معنای علّیت نیست؛ بنابراین حتی اگر correlation خیلی قوی نباشد، به این معنا نیست که رابطه‌ی علّی بین متغیرهای مستقل و وابسته وجود ندارد.

- ارتباط میان میزان فعالیت و PageRank:

برای هر شخص، تعداد آهنگ‌های منتشر شده توسط او را که با ستون num_release نمایش داده شده است، به عنوان میزان فعالیتش در نظر می‌گیریم.
برای بررسی قابل توجه بودن/نبودن ارتباط میان دو متغیر بیان شده، از مدل رگرسیون خطی و محاسبه‌ی correlation آن‌ها استفاده شد.
نتیجه‌ی مربوط به مدل رگرسیون خطی:

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.660e-03 -3.805e-05 -1.528e-05  2.175e-05  2.371e-03

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.040e-05  9.701e-07   51.96   <2e-16 ***
num_release 1.728e-06  2.470e-08   69.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000101 on 12903 degrees of freedom
Multiple R-squared:  0.275,     Adjusted R-squared:  0.2749
F-statistic:  4894 on 1 and 12903 DF,  p-value: < 2.2e-16
```

تیجه‌ی حاصل نشان می‌دهد که بین PageRank به عنوان متغیر وابسته و میزان فعالیت به عنوان متغیر مستقل، با مقدار t-value برابر ۶۹.۹۵ و p-value بسیار کوچک رابطه‌ی قابل توجه و معناداری با در نظر گرفتن سطح significance برابر ۰.۰۵ وجود دارد. این بدان معناست که به ازای هر افزایش یک واحدی در میزان فعالیت انتظار می‌رود مقدار PageRank به اندازه‌ی ۰.۰۰۰۰۰۱۷۲ واحد افزایش یابد.
significant بودن مدل به صورت کلی، با استفاده از مقدار آماره‌ی F تعیین می‌شود؛ که مقدار آن، این را نشان می‌دهد که آیا برازش کلی مدل، به طور قابل توجه و معناداری بهتر از حالت شانسی است یا خیر. حال، با توجه به اینکه مقدار آماره‌ی F برابر ۴۸۹۴ است و مقدار p-value بسیار کوچک است، قابل توجه و معنادار بودن عملکرد کلی مدل، نشان داده شده‌است. همچنین، مقدار R-squared تعدیل شده برابر ۰.۲۷۵ است و این را نشان می‌دهد که مدل تولید شده، تقریباً ۲۷.۵ درصد از واریانس در مقدار PageRank را توضیح می‌دهد، که میزان کمی‌ست.

نتیجه‌ی مربوط به محاسبه‌ی correlation بین این دو متغیر:

```
Correlation between pagerank and amount of activity:  0.5243757
```

مقدار متوسط رو به بالا محاسبه شده، وجود مقدار همبستگی نسبتاً بالایی میان این دو متغیر را نشان می‌دهد.

به طور کلی، نتایج حاصل از این دو تست آماری انجام شده را می‌توان بدین شرح تفسیر کرد:
با توجه به این‌که رگرسیون خطی وجود روابط قابل توجه و معناداری را نشان می‌دهد، و همچنین مقدار correlation مربوط به آن‌ها نسبتاً خوب است، این مسئله برداشت می‌شود که بین این دو متغیر، از نظر آماری رابطه‌ی معنادار و قابل توجهی وجود دارد؛ و این رابطه، نسبتاً قوی است.

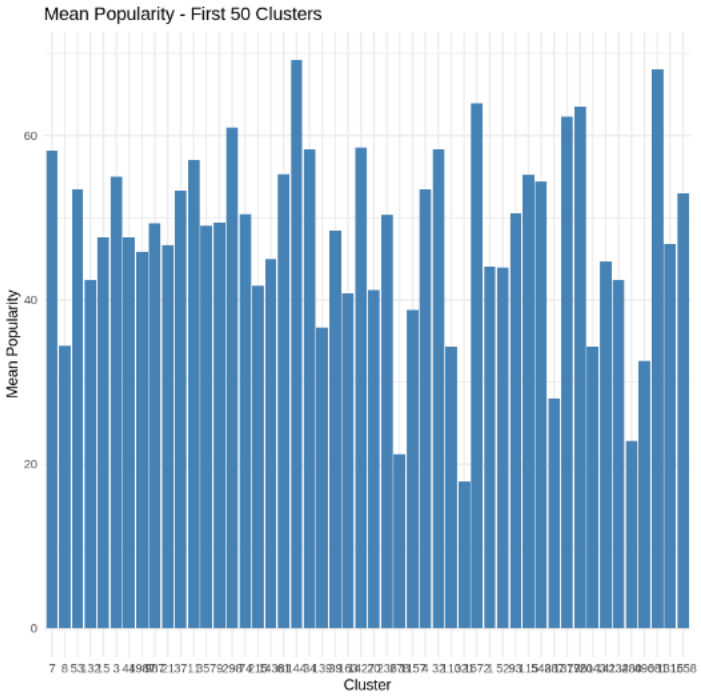
۶-۴ زیربخش ۶

الگوریتم label propagation یک الگوریتم برای یافتن community ها در یک گراف است. این الگوریتم، community ها را تنها با استفاده از ساختار شبکه شناسایی می‌کند و به یک تابع هدف از پیش تعریف شده یا به اطلاعات قبلی در مورد com-munity ها نیاز ندارد. این الگوریتم با انتشار برچسب‌ها در سراسر شبکه و تشکیل community ها بر اساس آن، کار می‌کند. در این فرایند، یک برچسب واحد می‌تواند به سرعت در یک گروه از گره‌ها منتشر شود، اما در عبور از یک منطقه پراکنده از گره‌ها، مشکل خواهد داشت. گره‌هایی که در پایان الگوریتم، دارای برچسب یکسان هستند را می‌توان بخشی از یک community در نظر گرفت.
برای خوشه بندی گره‌های گراف، از این الگوریتم استفاده شد. و عدد خوشه‌ی مربوط به هر یک از گره‌ها، در ستون جدیدی با نام cluster_label به دیتاست مربوط به گره‌ها اضافه شد.

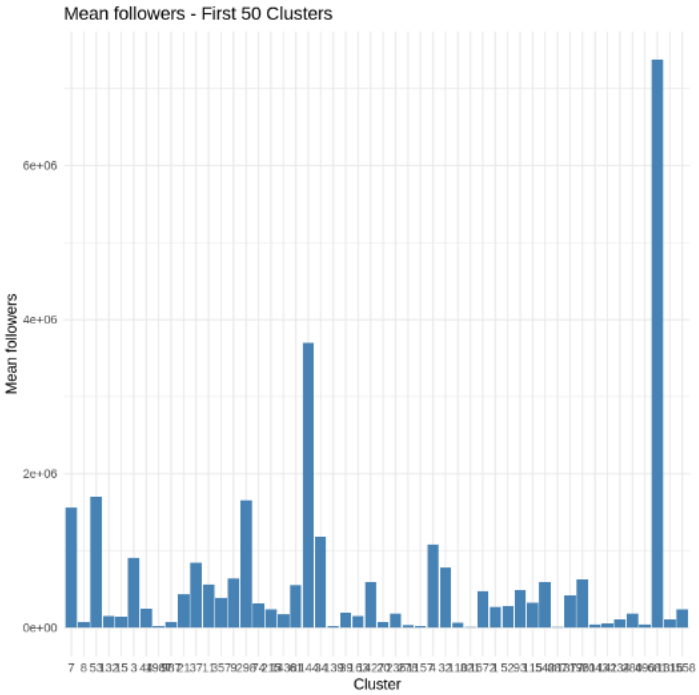
۷-۴ زیربخش ۷

دو خوشه با برچسب‌های ۷ و ۸، به ترتیب دو بزرگ‌ترین خوشه هستند.
برای ۵۰ خوشه‌ی دارای بیشترین سایز، به ترتیب نزولی سایز آن‌ها، نمودار ستونی مربوط به میانگین تعدادی از متغیرها رسم شده‌است:

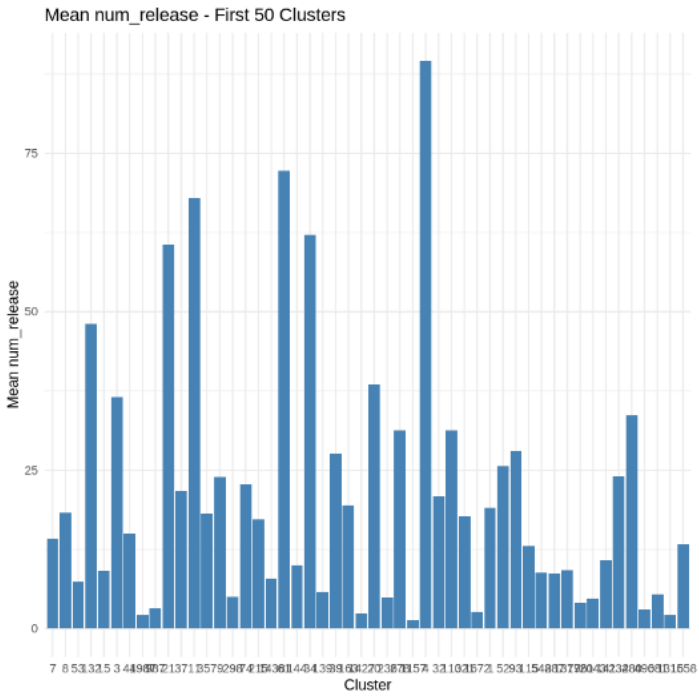
popularity -



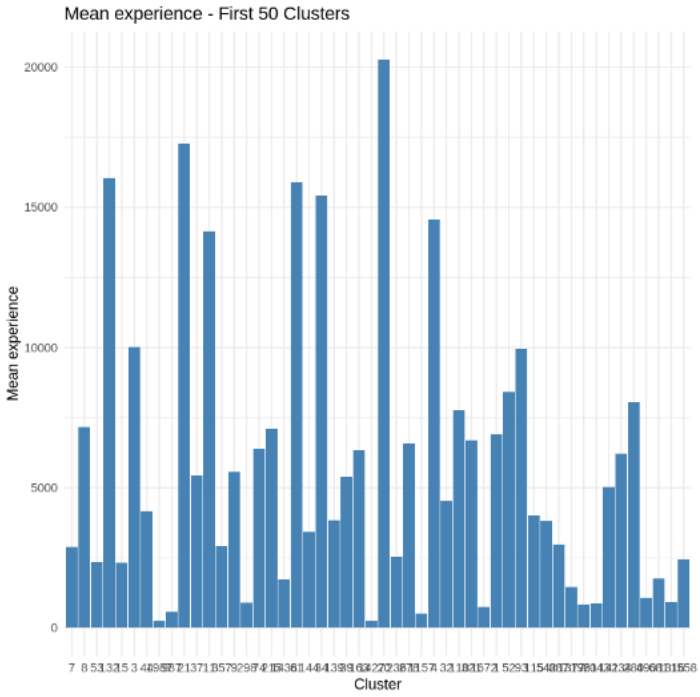
followers -



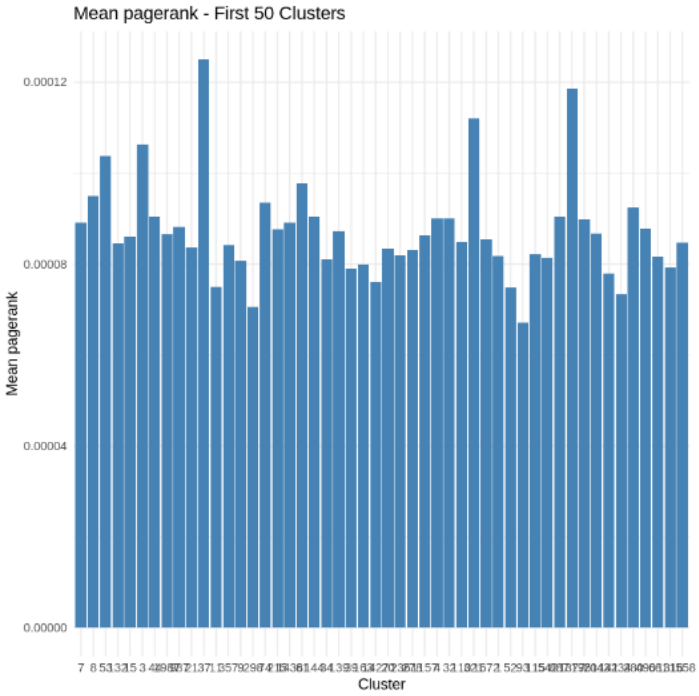
num_release -



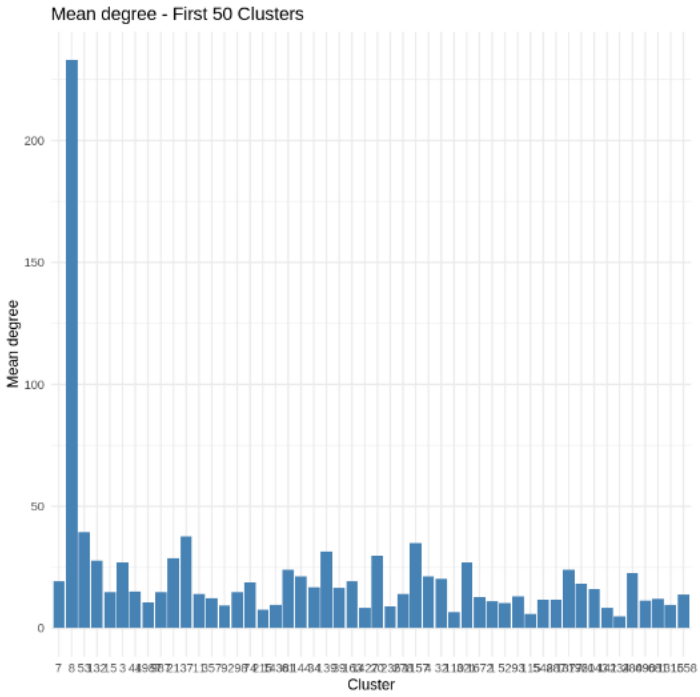
experience -



PageRank -



degree -



در هیچ یک از متغیرها، الگوی خاصی از جهت صعودی یا نزولی بودن، متناسب با سایز خوشه‌ها مشاهده نشد.

برای دو خوشه‌ی با بیشترین سایز، تعدادی فرض بر روش متغیرهای آن‌ها د رنظر می‌گیریم و سپس با استفاده از آزمون‌های مناسب، برقرار بودن/نبودن آن‌ها را بررسی می‌کنیم:

۱.

فرض صفر و فرض جایگزین:

-فرض صفر، این را بیان می‌کند که میانگین popularity برای دو خوشه یکسان است. و تفاوت significant ای (قابل توجهی) میان این میانگین‌ها وجود ندارد.

$$H_0 : \mu_1 = \mu_2 \quad (۴۷)$$

-فرض جایگزین، این را بیان می‌کند که میانگین popularity برای دو خوشه، یکسان نیست. و تفاوت significant ای (قابل توجهی) با هم دارند.

$$H_A : \mu_1 \neq \mu_2 \quad (۴۸)$$

برای بررسی این فرضیه، از آزمون T استفاده شد و نتایج به شرح زیر نشان می‌دهد که تفاوت قابل توجهی میان میانگبن popularity این دو خوشه وجود دارد:

```
Popularity:
AVG Popularity of cluster 1:
58.04531
AVG Popularity of cluster 2:
34.57771
There is a significant difference in average popularity between Cluster 1 and Cluster 2.
Test Statistic: 48.17838
P-Value: 0
```

۲.

فرض صفر و فرض جایگزین:

-فرض صفر، این را بیان می‌کند که میانگین followers برای دو خوشه یکسان است. و تفاوت significant ای (قابل توجهی) میان این میانگین‌ها وجود ندارد.

$$H_0 : \mu_1 = \mu_2 \quad (۴۹)$$

-فرض جایگزین، این را بیان می‌کند که میانگین followers برای دو خوشه، یکسان نیست. و تفاوت significant ای (قابل توجهی) با هم دارند.

$$H_A : \mu_1 \neq \mu_2 \quad (۵۰)$$

برای بررسی این فرضیه، از آزمون T استفاده شد و نتایج به شرح زیر نشان می‌دهد که تفاوت قابل توجهی میان میانگین followers این دو خوشه وجود دارد:

```
followers:
AVG followers of cluster 1:
1525626
AVG followers of cluster 2:
84135.54
There is a significant difference in average followers between Cluster 1 and Cluster 2.
Test Statistic: 14.66748
P-Value: 5.292299e-47
```

.۳

فرض صفر و فرض جایگزین:

-فرض صفر، این را بیان می‌کند که میانگین num_release برای دو خوشه یکسان است. و تفاوت significant ای (قابل توجهی) میان این میانگین‌ها وجود ندارد.

$$H_0 : \mu_1 = \mu_2 \quad (۵۱)$$

-فرض جایگزین، این را بیان می‌کند که میانگین num_release برای دو خوشه، یکسان نیست. و تفاوت significant ای (قابل توجهی) با هم دارند.

$$H_A : \mu_1 \neq \mu_2 \quad (۵۲)$$

برای بررسی این فرضیه، از آزمون T استفاده شد و نتایج به شرح زیر نشان می‌دهد که تفاوت قابل توجهی میان میانگین num_release این دو خوشه وجود دارد:

```
num_release:
AVG num_release of cluster 1:
13.92437
AVG num_release of cluster 2:
18.65066
There is a significant difference in average num_release between Cluster 1 and Cluster 2.
Test Statistic: -4.043703
P-Value: 5.394066e-05
```

.۴

فرض صفر و فرض جایگزین:

-فرض صفر، این را بیان می‌کند که میانگین experience برای دو خوشه یکسان است. و تفاوت significant ای (قابل توجهی) میان این میانگین‌ها وجود ندارد.

$$H_0 : \mu_1 = \mu_2 \quad (۵۳)$$

-فرض جایگزین، این را بیان می‌کند که میانگین experience برای دو خوشه، یکسان نیست. و تفاوت significant ای (قابل توجهی) با هم دارند.

$$H_A : \mu_1 \neq \mu_2 \quad (۵۴)$$

برای بررسی این فرضیه، از آزمون T استفاده شد و نتایج به شرح زیر نشان می‌دهد که تفاوت قابل توجهی میان میانگبن experience این دو خوشه وجود دارد:

```
experience:
AVG experience of cluster 1:
2810.24
AVG experience of cluster 2:
7321.052
There is a significant difference in average experience between Cluster 1 and Cluster 2.
Test Statistic: -25.06699
P-Value: 9.245606e-126
```

نتیجه‌ی این بررسی‌ها نشان می‌دهد که خوشه‌ی اول از نظر میانگین امتیاز محبوبیت و میانگین تعداد دنبال کننده‌ها، نسبت به خوشه‌ی دوم برتری دارد. و خوشه‌ی دوم نیز از نظر میانگین تعداد آهنگ‌های منتشرشده و میانگین زمان تجربه، نسبت به خوشه‌ی اول برتری دارد.

۸-۴ زیربخش ۸

ابتدا گره‌های مربوط به هر یک از ۳ ژانر بیان شده، در ۳ دیتاست قرار گرفت. سپس، میانگین متغیرهای خواسته شده برای هر یک از این دیتاست‌ها به شرح زیر محاسبه شد:

- میانگین تعداد آهنگ‌ها

```
Average number of songs - Pop: 23.16842
Average number of songs - Hip Hop: 31.74
Average number of songs - Classical: 71.95789
```

- میانگین سال‌های تجربه

```
Average years of experience - Pop: 2800.813
Average years of experience - Hip Hop: 5845.707
Average years of experience - Classical: 17615.65
```

- میانگین تعداد همکاری

```
Average number of collaborations - Pop: 37.33053
Average number of collaborations - Hip Hop: 41.66
Average number of collaborations - Classical: 1212.968
```

- میانگین تعداد دنبال کنندگان

```
Average number of followers - Pop: 4992242
Average number of followers - Hip Hop: 2300181
Average number of followers - Classical: 201959.7
```

- میانگبن امتیاز محبوبیت

```
Average number of popularity - Pop: 74.00842
Average number of popularity - Hip Hop: 67.54
Average number of popularity - Classical: 48.93158
```

همانطور که مشاهده می‌شود، خوانندگان مربوط به ژانر classical دارای بیشترین میانگین تعداد آهنگ‌ها، بیشترین میانگین سال‌های تجربه و بیشترین میانگین تعداد همکاری هستند. اما با این حال، دارای کمترین میانگین تعداد دنبال کنندگان و کمترین میانگین امتیاز محبوبیت هستند. در نتیجه، مشهور شدن در این ژانر، سخت‌تر است.

۹-۴ زیربخش ۹

فرض صفر و فرض جایگزین:
- یال‌های این گراف و گراف رندوم از نظر pattern ها یکسانند و دارای تفاوت قابل توجهی نیستند.
- یال‌های این گراف و گراف رندوم از نظر pattern ها یکسان نیستند و دارای تفاوت قابل توجهی هستند.
ابتدا مقدار ضریب خوشه بندی (معیار استفاده شده برای بررسی pattren یال‌های گراف) برای شبکه‌ی در دسترس محاسبه شد. سپس تعدادی گراف رندوم دارای تعداد گره‌ها و چگالی یال‌های برابر با گراف در دسترس، ساخته شد. برای هر یک از این گراف‌ها مقدار ضریب خوشه بندی محاسبه شد. در نهایت بین این ضریب خوشه بندی‌هایی که برای گراف‌های رندوم محاسبه شد و ضریب خوشه بندی ای که برای شبکه‌ی در دسترس محاسبه شده بود، p-value محاسبه شد. p-value برابر با ۰.۰۸ بدست آمد، که بیشتر از سطح significance برابر ۰.۰۵ است و فرض صفر را نمی‌توان رد کرد و نشان‌دهنده‌ی این است که یال‌های این گراف و گراف رندوم از نظر pattern ها یکسانند و دارای تفاوت قابل توجهی نیستند.