EXPERIMENT - 4

AIM: Implementation of Statistical Hypothesis Test using Scipy and Sci-kit learn

TO - DO:

Correlation Tests:
1. Pearson's Correlation Coefficient:
2. Spearman's Rank Correlation
3. Kendall's Rank Correlation:
4. Chi-Squared Test

ABOUT DATASET:

Link to our dataset:
https://www.kaggle.com/datasets/bartomiejczyewski/spotify-top-200-weekly-global-20172021

This dataset includes Global Top 200 songs weekly from Spotify. Charts were scrapped from: https://spotifycharts.com/regional, additional information about the songs was taken from the Spotify API. This dataset can be interesting both for beginners and advanced Data Scientists

The columns are:- Rank, Track,Artist, stream,link,album name,duration_MS,Explicit,Track_number_on_album,Artist_Followers,Artist_Genres

THEORY:

1.Pearson's Correlation Coefficient:

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

2. Spearman's Rank Correlation :

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

The formula for Spearman's rank coefficient is:

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2-1)}$$

ρ = Spearman's rank correlation coefficient

di = Difference between the two ranks of each observation

n = Number of observations

The Spearman Rank Correlation can take a value from +1 to -1 where,

- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

3. Kendall's Rank Correlation:

Kendall rank correlation (non-parametric) is an alternative to Pearson's correlation (parametric) when the data you're working with has failed one or more assumptions of the test. This is also the best alternative to Spearman correlation (non-parametric) when your sample size is small and has many tied ranks.

Kendall rank correlation is used to test the similarities in the ordering of data when it is ranked by quantities. Other types of correlation coefficients use the observations as the basis of the correlation, Kendall's correlation coefficient uses pairs of observations and determines the strength of association based on the patter on concordance and discordance between the pairs.

- **Concordant:** Ordered in the same way (consistency). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have the same sign.
- **Discordant:** Ordered differently (inconsistency). A pair of observations is considered concordant if (x2 — x1) and (y2 — y1) have opposite signs.

Kendall's Tau coefficient of correlation is usually smaller than Spearman's rho correlation. The calculations are based on concordant and discordant pairs. Insensitive to error. P values are more accurate with smaller sample sizes.

4. Chi-Squared Test:

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences.

It is used to calculate the difference between two categorical variables, which are:

- As a result of chance or
- Because of the relationship

Formula For Chi-Square Test:

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Where
c = Degrees of freedom
O = Observed Value
E = Expected Value

SCREENSHOTS:

- **Loading the dataset**

```
[1] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sb
```

```
#loading the dataset
data = pd.read_csv("Spotify Top 200 Global (2017-2021).csv")
data
```

0	1	Starboy	The Weeknd	25734078	https://open.spotify.com/track/5aAx2yezTd8zXrk...	2017-01-06	Starboy	230453	True	1
1	2	Closer	The Chainsmokers	23519705	https://open.spotify.com/track/7BKLCZ1jbUBVqRi...	2017-01-06	Closer	244960	False	1
2	3	Rockabye (feat. Sean Paul & Anne-Marie)	Clean Bandit	21216399	https://open.spotify.com/track/5knuzwU65gJK7IF...	2017-01-06	Rockabye (feat. Sean Paul & Anne-Marie)	251088	False	1
3	4	Let Me Love You	DJ Snake	19852704	https://open.spotify.com/track/4pdPtRcBmOSQDlJ...	2017-01-06	Encore	205946	False	13
4	5	I Don't Wanna Live Forever (Fifty Shades Darke...	ZAYN	18316326	https://open.spotify.com/track/3NdDpSvN911VPGi...	2017-01-06	I Don't Wanna Live Forever (Fifty Shades Darker)	245200	False	1
...	...	...	...	...	...	...	...	...	...	...
44195	196	Can't Hold Us - feat. Ray	Macklemore & Ryan Lewis	5174246	https://open.spotify.com/track/3bidbhpOYeV4knp...	2021-04-16	The Heist	258342	False	2

Executing (2m 22s) Cell > inner_f() > heatmap() > plot() > _annotate_heatmap() > relative_luminance()

- **Pearson's Correlation Coefficient**

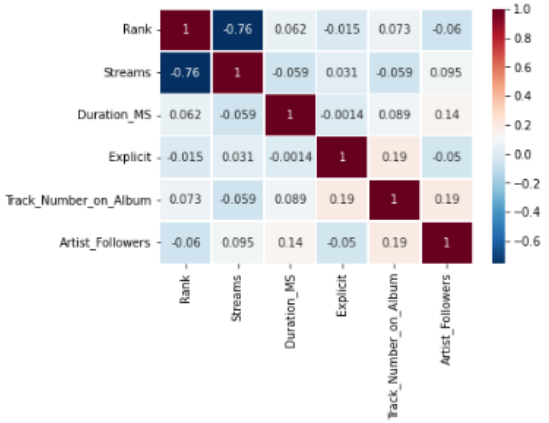
```
[3] d1 = data.corr(method='pearson')
    d1
```

	Rank	Streams	Duration_MS	Explicit	Track_Number_on_Album	Artist_Followers
Rank	1.000000	-0.757160	0.061727	-0.015431	0.073173	-0.059790
Streams	-0.757160	1.000000	-0.059353	0.031118	-0.059379	0.095335
Duration_MS	0.061727	-0.059353	1.000000	-0.001396	0.088745	0.140047
Explicit	-0.015431	0.031118	-0.001396	1.000000	0.190453	-0.049625
Track_Number_on_Album	0.073173	-0.059379	0.088745	0.190453	1.000000	0.190726
Artist_Followers	-0.059790	0.095335	0.140047	-0.049625	0.190726	1.000000

```
[4] import seaborn as sb
```

```
sb.heatmap(d1,
           xticklabels=d1.columns,
           yticklabels=d1.columns,
           cmap='RdBu_r',
           annot=True,
           linewidth=0.5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9f0affe820>
```



- **Spearman's Rank Correlation**

Spearman's rank order correlation

```
[6] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    from scipy.stats import spearmanr
```

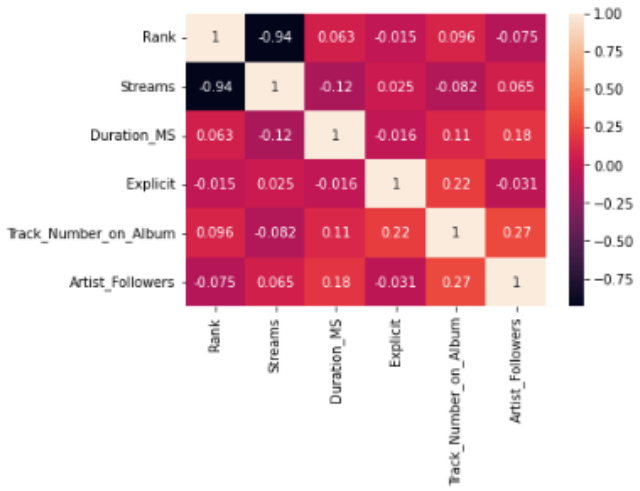
```
#Calculate the Spearman's rank correlation for a pair of columns
data['Artist_Followers'].corr(data['Streams'],method='spearman')
```

```
0.06471370120731422
```

```
[8] #Calculate the Spearman's rank correlation for all columns in the dataframe
    data.corr(method='spearman' )
```

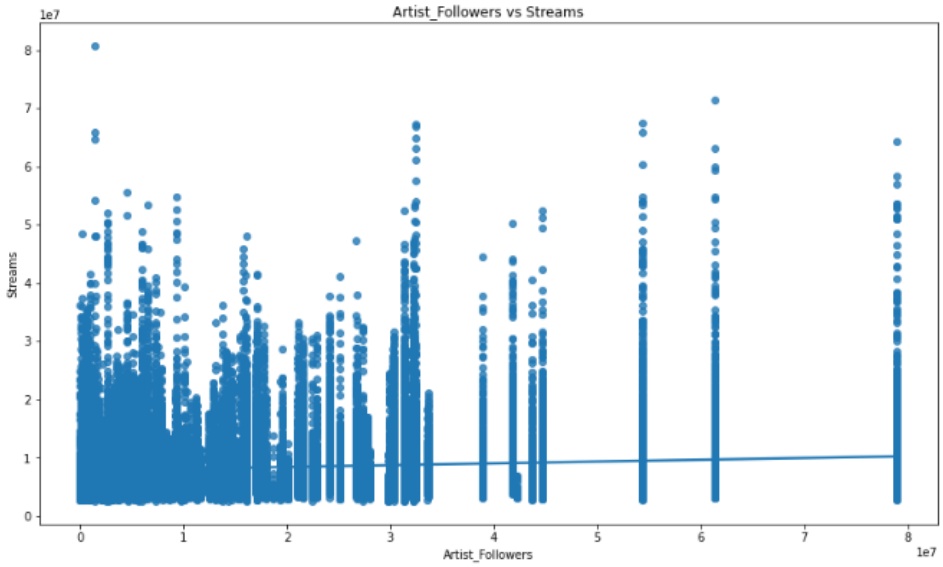
	Rank	Streams	Duration_MS	Explicit	Track_Number_on_Album	Artist_Followers
Rank	1.000000	-0.936006	0.062692	-0.015431	0.096158	-0.075472
Streams	-0.936006	1.000000	-0.115650	0.025439	-0.081789	0.064714
Duration_MS	0.062692	-0.115650	1.000000	-0.015539	0.110870	0.180662
Explicit	-0.015431	0.025439	-0.015539	1.000000	0.215787	-0.031457
Track_Number_on_Album	0.096158	-0.081789	0.110870	0.215787	1.000000	0.265033
Artist_Followers	-0.075472	0.064714	0.180662	-0.031457	0.265033	1.000000

```
#Visualise Spearman's rank correlation coefficients with a heatmap
corr = data.corr(method='spearman')
sns.heatmap(corr,annot=True)
plt.show()
```




```
#Visualising correlations between pairs of columns
plt.figure(figsize=(14,8))
sns.regplot(x='Artist_Followers', y='Streams', data=data)
plt.title('Artist_Followers vs Streams')
```

```
Text(0.5, 1.0, 'Artist_Followers vs Streams')
```



```
[11] #Calculate the statistical significance of the correlation coefficient
     spearmanr(data['Artist_Followers'], data['Streams'])
     #p value is more thus it is not statistically significant
```

```
SpearmanrResult(correlation=0.06471370120731422, pvalue=3.080569036257845e-42)
```

```
[12] spearmanr(data['Track_Number_on_Album'], data['Streams'])
     #p value is comparatively less thus it is statistically significant
```

```
SpearmanrResult(correlation=-0.08178890918724431, pvalue=1.7693319186043194e-66)
```

- **Kendall's Rank Correlation**

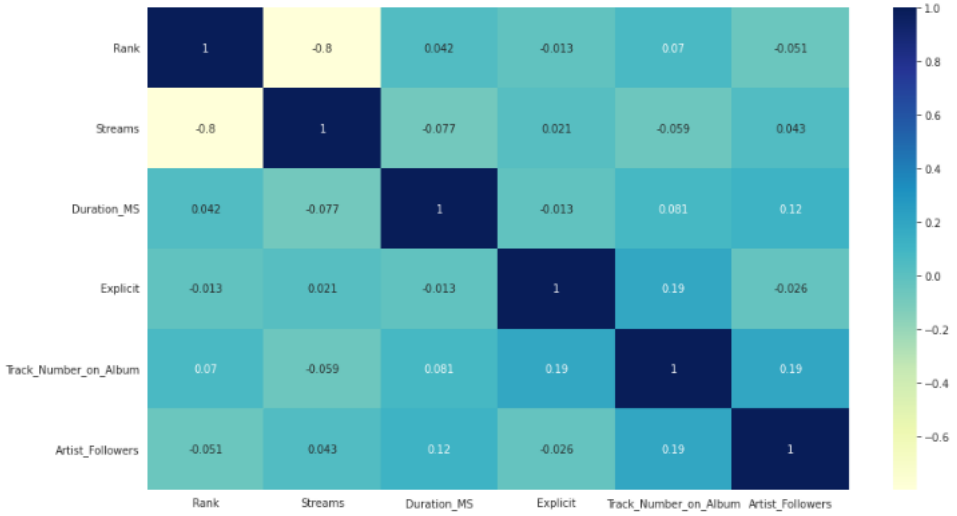
```
[13] #Kendall Rank
     import pandas as pd
     from pylab import rcParams
     import seaborn as sb
     from scipy.stats.stats import kendalltau
```

```
[14] # Data Visualisation Settings
     %matplotlib inline
     rcParams['figure.figsize'] = 5,4
     sb.set_style('whitegrid')
```

```
corr = data.corr(method='kendall')
```

```
rcParams['figure.figsize'] = 14.7,8.27
sb.heatmap(corr,
           xticklabels=corr.columns.values,
           yticklabels=corr.columns.values,
           cmap="YlGnBu",
           annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9f07f53fd0>
```



- **Chi-Squared Test**

```
#Null hypothesis - There is no relation between Artist followers and streams
#Alternative hypothesis - There is no relation between Artist followers and streams
```

+ Code + Text

```
import pandas as pd
from scipy.stats import chi2_contingency
import seaborn as sns
import matplotlib.pyplot as plt
path = '/content/Spotify Top 200 Global (2017-2021).csv'
spotifyData = pd.read_csv(path)
contingency= pd.crosstab(spotifyData['Streams'], spotifyData['Artist_Followers'])

c, p, dof, expected = chi2_contingency(contingency)
print(p)
```

```
0.04084312286693688
```

+ Code + Text

```
[19] #Since the p value is less than 0.5, we will reject the null hypothesis
     #Thus our alternative hypothesis will get accepted
     #Thus There is relation between Artist followers and streams
```

CONCLUSION: We performed the Pearson's correlation coefficient test and visualized it. After that we did the spearman's correlation coefficient test where we compared columns and visualized them. We visualized Kendall's rank correlation. After that, In the Chi square test, we got p value less than 0.5. Thus we accepted the alternative hypothesis.