DATA ANALYTICS (CS40003) PROJECT REPORT

(2019-20 Autumn)

Title of the Project:
Spearman's correlation analysis for paired data(ProjectID: 4)

Date of Submission: 06/12/2019

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Team Details:
Group Id: 64
Nikil Raghute(16AG36015)

Problem Statement:

Spearman's correlation analysis for paired data

Reference: **SNACKS** data

a) Find the Spearman correlation matrix of all the ordinal attributes.
b) Determine the coefficient of determination.
c) Interpret the result from the two tables.
d) In each case, perform the significance test with 95% confidence level.

Theory:

We'll need the concepts of Spearman's correlation coefficient, coefficient of Determination, significance test.

Correlation:

is used to donate some form of association between two variables, how strongly pairs of variables are related.

- $r = 0$, implies there is no correlation .
- $r = +1$ (perfect positive correlation).
- $r = -1$ (perfect negative correlation).
- Value of r nearer to **+1 or -1** indicates **high degree of correlation** between the two variables.

Charles Spearman's coefficient of correlation:

- Is used to find correlation coefficient between two **ordinal attributes.**
- This correlation measurement is also called **Rank correlation**.
- This technique is applicable to determine the degree of correlation between two variables in case of ordinal data.
- It assesses how well the relationship between two variables can be described using a monotonic function.

For example:

Let's consider we've two columns "Col 1" and "Col 2".

We can find rs as follows:

First we'll calculate rank in respective columns, taking differences of their ranks and summing the square of differences of their ranks.

Sl No.	Col 1	rank1	Col 2	rank2	d=rank1-rank2	d^2
1	x1	r1	y1	s1	r1-s1	(r1-s1)^2
2	x2	r2	y2	s2	r2-s2	(r2-s2)^2
3	x3	r3	y3	s3	r3-s3	(r3-s3)^2
4	x4	r4	y4	s4	r4-s4	(r4-s4)^2
5	x5	r5	y5	s5	r5-s5	(r5-s5)^2
6	x6	r6	y6	s6	r6-s6	(r6-s6)^2
7	x7	r7	y7	s7	r7-s7	(r7-s7)^2
8	x8	r8	y8	s8	r8-s8	(r8-s8)^2
						Σdi^2

After that, we can use the below formula,

Formula:

The rank correlation can be defined as

$$r_s = 1 - \frac{(6 * \sum di^2)}{n(n^2-1)}$$

Where, d_i=Difference between ranks of i^th pair of the two variables

n=Number of pairs of observations.

$-1 <= r_s <= 1$

Coefficient of Determination:

- It is used to measure the proportion of variability of the fitted model.
- It is square of correlation(r), thus varies between 0 and 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.

Significance Test:

We can carry out significance test in 5 steps:
Step 1- Defining a Hypothesis.

Step 2- Finding r_s (using their ranks).

Step 3- Finding r_s value from the Spearman's table/graph for given DOF and significance level.

Step 4- Verifying if calculated r_s is higher or lower than r_s from the table/graph.

Step 5- Rejecting(if calculated rs is higher) or Fail to reject the hypothesis.
And the final comment.

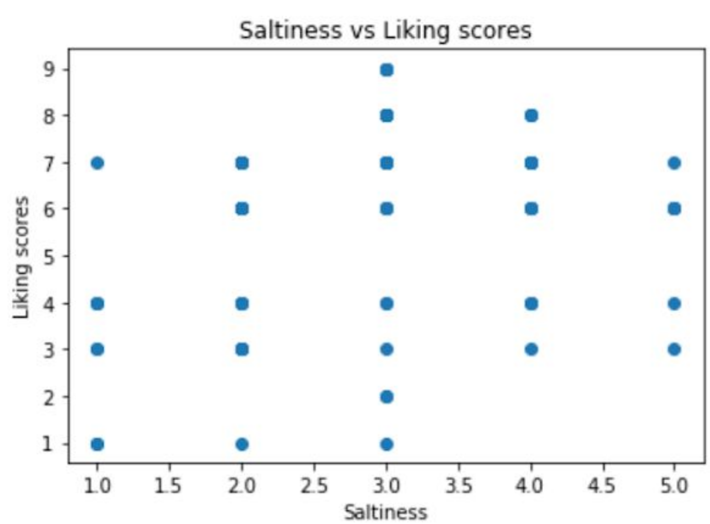
Calculations:

a) Spearman’s correlation coefficients-

Let’s calculate Spearman’s correlation coefficients($\mathbf{r_s}$) for our **“SNACKS”** data.
We’ve our **SNACKS** dataset stored in **“data” DataFrame**.
We can see it’s first 5 rows using:
data.head()

	Liking scores	Saltiness	Sweetness	Acidity	Crunchiness
0	1	3	3	3	3
1	1	1	2	3	1
2	1	2	2	5	1
3	1	1	4	3	1
4	2	3	3	2	2

We can try plotting feature vs target variable,
For example, the scatterplot of Saltiness vs Liking scores, all plots are very complex and we can’t say if there can be any correlation between them.



Now, let's calculate $\mathbf{r_s}$,

By using for loop, we can calculate $\mathbf{r_s}$ among all possible pairs of attributes,

	Liking scores	Saltiness	Sweetness	Acidity	Crunchiness
Liking scores	1.000000	0.319130	0.101599	0.031476	0.509529
Saltiness	0.319130	1.000000	-0.006757	0.022190	0.211491
Sweetness	0.101599	-0.006757	1.000000	-0.098605	0.049244
Acidity	0.031476	0.022190	-0.098605	1.000000	0.250573
Crunchiness	0.509529	0.211491	0.049244	0.250573	1.000000

But as we're only interested in $\mathbf{r_s}$ of pairs between features and target variable. We can use it from the above table, but for simplification and more details explanation, let's calculate separately one-by-one.

i) Saltiness- Liking scores

We can use "Saltiness" and "Liking scores" columns, computing their ranks and then following the procedure discussed in the theory part:

Here, printing first 5 rows of the table:

	Saltiness	Salt r	Liking scores	Liking scores r	diff	diff2
0	3	46.0	1	98.5	-52.5	2756.25
1	1	96.5	1	98.5	-2.0	4.00
2	2	79.5	1	98.5	-19.0	361.00
3	1	96.5	1	98.5	-2.0	4.00
4	3	46.0	2	95.5	-49.5	2450.25

So we've got, d^2 = 2756.25 + 4.00 + 361.00 + 4.00 + 2450.25 = 113467.0

n = 100

Therefore,

$$\mathbf{r_s} = 1 - \frac{(6 * \sum di^2)}{n(n^2-1)} = \mathbf{1} - \frac{6*2450.25}{100(100^2-1)} = \mathbf{0.3191299 \approx 0.319}$$

Similarly, we can calculate for all other pairs.

ii) Sweetness- Liking scores

First five rows of the table:

	Sweetness	Sweet r	Liking scores	Liking scores r	diff	diff2
0	3	56.5	1	98.5	-42.0	1764.00
1	2	86.0	1	98.5	-12.5	156.25
2	2	86.0	1	98.5	-12.5	156.25
3	4	21.0	1	98.5	-77.5	6006.25
4	3	56.5	2	95.5	-39.0	1521.00

d^2 = 149718.5

n = 100

r_s = 0.101599159915 ≈ 0.102

iii) Acidity- Liking scores

First five rows of the table:

	Acidity	Acid r	Liking scores	Liking scores r	diff	diff2
0	3	53.0	1	98.5	-45.5	2070.25
1	3	53.0	1	98.5	-45.5	2070.25
2	5	5.0	1	98.5	-93.5	8742.25
3	3	53.0	1	98.5	-45.5	2070.25
4	2	85.5	2	95.5	-10.0	100.00

d^2 = 161404.5

n = 100

r_s = 0.0314761476 ≈ 0.031

iv) Crunchiness- Liking scores

First five rows of the table:

	Crunchiness	Crunch r	Liking scores	Liking scores r	diff	diff2
0	3	37.0	1	98.5	-61.5	3782.25
1	1	96.5	1	98.5	-2.0	4.00
2	1	96.5	1	98.5	-2.0	4.00
3	1	96.5	1	98.5	-2.0	4.00
4	2	76.0	2	95.5	-19.5	380.25

$d^2 = 81737.0$

n = 100

r_s = 0.5095289528 ≈ 0.509

So finally we've **Spearman's correlation coefficient(r_s)** for different pairs as given below:

Pairs	rs
Saltiness- Liking scores	0.319
Sweetness- Liking scores	0.102
Acidity- Liking scores	0.031
Crunchiness- Liking scores	0.509

b) Coefficient of Determination(R^2):

The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor.

Since $R^2 = r_s * r_s$

Coefficient of Determination for:

- Saltiness- Liking scores = 0.319*0.319 = **0.102**
- Sweetness- Liking scores = 0.102*0.102 = **0.010**
- Acidity- Liking scores = 0.031*0.031= **0.00099**
- Crunchiness- Liking scores = 0.509*0.509 = **0.259**

Pairs	Coefficient of Determination(R^2)
Saltiness- Liking scores	0.102
Sweetness- Liking scores	0.01
Acidity- Liking scores	0.00099
Crunchiness- Liking scores	0.259

c) Interpreting the results from (a) and (b):

From (a), from calculated values of r_s we can say that **"Saltiness"** and **"Crunchiness"** are **fairly rank-correlated**(fair monotonic relation) to "Liking scores", while "Sweetness" and "Acidity" are not significantly correlated.

From (b), as we've calculated **Coefficient of Determination(** R^2 **)**, we can say that **level of variance** in the dependent variable caused by its relationship with the independent variable **is higher** in case of **"Saltiness"** and **"Crunchiness"** compared to "Sweetness" and "Acidity".

d) Significance Test:

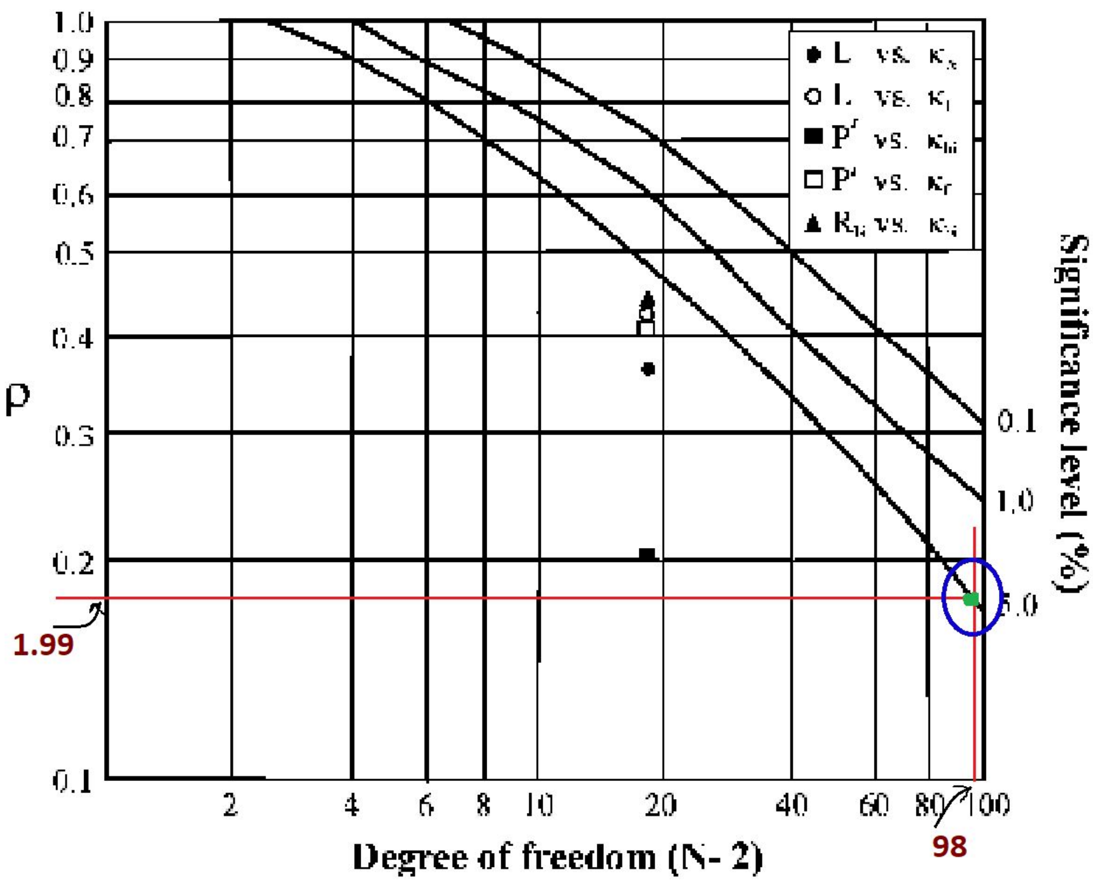
We can use the Spearman's coefficient as statistical method for proving or disproving a hypothesis.

C.I = 95%
So **α = 5% = 0.05**(Two-tail test)

Hypothesis:

H0: The variables **do not** have a **rank-order relationship** in the data.
To reject H0: is to say that there is a **rank-order relationship** between the variables in the data.

N = 100
Degree of Freedom(DOF) = 100-2 = 98
α = 0.05



Critical Values of the Spearman's Ranked Correlation Coefficient (r_s)
Taken from Zar, 1984 Table B.19

α(2):	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
α(1):	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
n									
4	0.600	1.000	1.000						
5	0.500	0.800	0.900	1.000	1.000				
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000		
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12	0.217	0.406	0.503	0.587	0.678	0.727	0.769	0.818	0.846
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.791	0.824
14	0.200	0.367	0.464	0.538	0.626	0.679	0.723	0.771	0.802
15	0.189	0.354	0.446	0.521	0.604	0.654	0.700	0.750	0.779
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.729	0.762
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37	0.114	0.216	0.275	0.325	0.383	0.421	0.456	0.497	0.526
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507
41	0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501
42	0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495
43	0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490
44	0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484
45	0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479
46	0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474
47	0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469
48	0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465
49	0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460
50	0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456
51	0.096	0.182	0.233	0.276	0.326	0.359	0.390	0.426	0.451
52	0.095	0.180	0.231	0.274	0.323	0.356	0.386	0.422	0.447
53	0.095	0.179	0.228	0.271	0.320	0.352	0.382	0.418	0.443
54	0.094	0.177	0.226	0.268	0.317	0.349	0.379	0.414	0.439
55	0.093	0.175	0.224	0.266	0.314	0.346	0.375	0.411	0.435
56	0.092	0.174	0.222	0.264	0.311	0.343	0.372	0.407	0.432
57	0.091	0.172	0.220	0.261	0.308	0.340	0.369	0.404	0.428
58	0.090	0.171	0.218	0.259	0.306	0.337	0.366	0.400	0.424
59	0.089	0.169	0.216	0.257	0.303	0.334	0.363	0.397	0.421
60	0.089	0.168	0.214	0.255	0.300	0.331	0.360	0.394	0.418
61	0.088	0.166	0.213	0.252	0.298	0.329	0.357	0.391	0.414
62	0.087	0.165	0.211	0.250	0.296	0.326	0.354	0.388	0.411
63	0.086	0.163	0.209	0.248	0.293	0.323	0.351	0.385	0.408
64	0.086	0.162	0.207	0.246	0.291	0.321	0.348	0.382	0.405
65	0.085	0.161	0.206	0.244	0.289	0.318	0.346	0.379	0.402
66	0.084	0.160	0.204	0.243	0.287	0.316	0.343	0.376	0.399
67	0.084	0.158	0.203	0.241	0.284	0.314	0.341	0.373	0.396
68	0.083	0.157	0.201	0.239	0.282	0.311	0.338	0.370	0.393
69	0.082	0.156	0.200	0.237	0.280	0.309	0.336	0.368	0.390
70	0.082	0.155	0.198	0.235	0.278	0.307	0.333	0.365	0.388
71	0.081	0.154	0.197	0.234	0.276	0.305	0.331	0.363	0.385
72	0.081	0.153	0.195	0.232	0.274	0.303	0.329	0.360	0.382
73	0.080	0.152	0.194	0.230	0.272	0.301	0.327	0.358	0.380
74	0.080	0.151	0.193	0.229	0.271	0.299	0.324	0.355	0.377
75	0.079	0.150	0.191	0.227	0.269	0.297	0.322	0.353	0.375
76	0.078	0.149	0.190	0.226	0.267	0.295	0.320	0.351	0.372
77	0.078	0.148	0.189	0.224	0.265	0.293	0.318	0.349	0.370
78	0.077	0.147	0.188	0.223	0.264	0.291	0.316	0.346	0.368
79	0.077	0.146	0.186	0.221	0.262	0.289	0.314	0.344	0.365
80	0.076	0.145	0.185	0.220	0.260	0.287	0.312	0.342	0.363
81	0.076	0.144	0.184	0.219	0.259	0.285	0.310	0.340	0.361
82	0.075	0.143	0.183	0.217	0.257	0.284	0.308	0.338	0.359
83	0.075	0.142	0.182	0.216	0.255	0.282	0.306	0.336	0.357
84	0.074	0.141	0.181	0.215	0.254	0.280	0.305	0.334	0.355
85	0.074	0.140	0.180	0.213	0.252	0.279	0.303	0.332	0.353
86	0.074	0.139	0.179	0.212	0.251	0.277	0.301	0.330	0.351
87	0.073	0.139	0.177	0.211	0.250	0.276	0.299	0.328	0.349
88	0.073	0.138	0.176	0.210	0.248	0.274	0.298	0.327	0.347
89	0.072	0.137	0.175	0.209	0.247	0.272	0.296	0.325	0.345
90	0.072	0.136	0.174	0.207	0.245	0.271	0.294	0.323	0.343
91	0.072	0.135	0.173	0.206	0.244	0.269	0.293	0.321	0.341
92	0.071	0.135	0.173	0.205	0.243	0.268	0.291	0.319	0.339
93	0.071	0.134	0.172	0.204	0.241	0.267	0.290	0.318	0.338
94	0.070	0.133	0.171	0.203	0.240	0.265	0.288	0.316	0.336
95	0.070	0.133	0.170	0.202	0.239	0.264	0.287	0.314	0.334
96	0.070	[illegible]	0.169	0.201	0.238	0.262	0.285	0.313	0.332
97	0.069	0.131	[illegible]	0.200	0.236	0.261	0.284	0.311	0.331
98	0.069	0.130	0.167	0.199	0.235	0.260	0.282	0.310	0.329
99	0.068	0.130	0.166	0.198	0.234	0.258	0.281	0.308	0.327
100	0.068	0.129	0.165	[illegible]	0.233	0.257	0.279	0.307	0.326

0.199

From the **Spearman's rank correlation coefficient graph and table** we can find **Spearman's coefficient as <u>0.199.</u>**

i) Saltiness- Liking scores

rs = 0.319

and Spearman's rank correlation coefficient from Spearman's rank significance table is 0.199.

as **0.319 > 0.199, we reject the hypothesis**, i.e. there is a greater than 95% chance that the **relationship is significant(not random)** among "Saltiness" and "Liking scores" attributes.

ii) Sweetness- Liking scores

rs = 0.102
And from the table **0.199,**

As **0.102 < 0.199**, we **fail to reject** the hypothesis, the variables **do not have** a significant rank-order relationship in the data.

iii) Acidity- Liking scores

rs = 0.031
And from the table **0.199,**

As **0.031 < 0.199,** we **fail to reject** the hypothesis, the variables **do not have** a significant rank-order relationship in the data.

iv) Crunchiness- Liking scores

rs = 0.509
And from the table **0.199**

As **0.509 > 0.199,** we **reject the** hypothesis, i.e. i.e. **the relationship is significant(not random)**.

Experimental Results:

- As r_s values for **Sweetness-Liking scores** and **Acidity-Liking scores** are very **low(nearer to 0)**, we can conclude that Sweetness and Liking scores are **not much rank-correlated**, same for the Acidity- Liking scores pair.

- As r_s values for **Saltiness and Liking scores**, **Crunchiness and Liking scores** are **nearer to 0.5**, attributes are **fairly rank-correlated**.

- As r_s value for **Crunchiness and liking scores** is relatively higher, these variables are more correlated any than others.

- Similarly, measure of variability of one factor can be caused by its relationship to another factor is in the order:
  Crunchiness-Liking scores > Saltiness-Liking scores > Sweetness-Liking scores > Acidity-Liking scores.
  i.e. “Liking scores” can be calculated from “Crunchiness” with less error compare to other features.

- **From Significance test:**

  i)We **reject** the hypothesis in **Saltiness-Liking scores** and **Crunchiness-Liking scores**. We can conclude that, there is a **significant relationship(i.e not random)** between Saltiness & Liking scores, and between Crunchiness & Liking scores.

  ii) In **Sweetness-Liking scores** and **Acidity-Liking scores**, we **fail to reject** hypothesis(H_0) and conclude that “Sweetness” and “Liking scores” **do not have a significant rank-order relationship**, same in the case of “Acidity” and “Liking scores”.

THANK YOU!