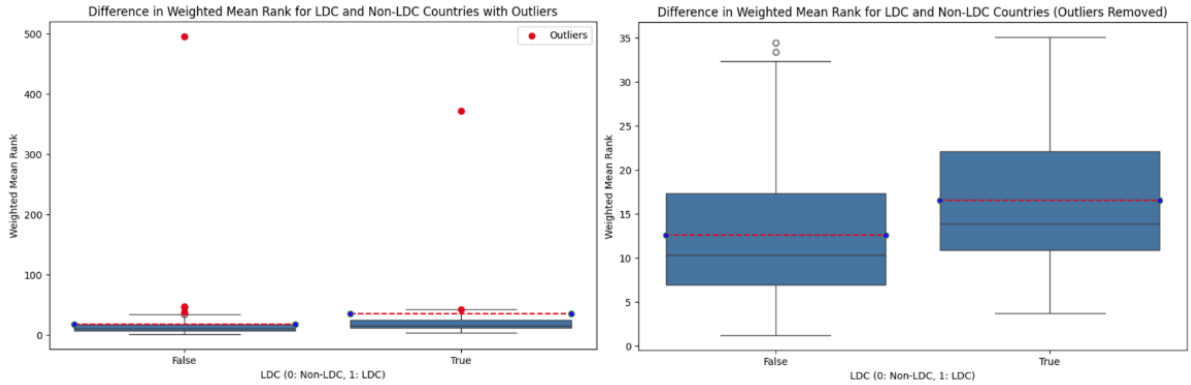
ROCS-M Coursework

1. (I) To calculate Instagrams popularity across the worlds Least Developed Countries (LDCs), I used the formula below for each country. This metric adjusts for variations in data availability by accounting for the number of days Instagram appeared in the rankings relative to the total days covered per month by the dataset (31). This approach makes sure that countries with incomplete ranking data are not overrepresented, resulting in a more accurate and balanced measure of popularity. I implemented this formula by filtering Instagram-specific data, grouping it by country, and applying the formula to calculate the 'weighted_mean_rank' for each country. 'Weighted_mean_rank' therefore represents popularity.

$$\text{Weighted Mean Rank} = \frac{\text{Total Days in Dataset}}{\text{Days Instagram Appeared}} \text{ x Average Rank}$$

(II) (a) To ensure the visual was not skewed by extreme values, I removed outliers creating a more accurate representation of the data. Additionally, I added a red dashed line to show the mean for each group. This made it easier to compare the mean vs the median and provides a clearer understanding of the overall trend.



In the first box-plot (left), before removing the outliers, we can see some extreme values in both LDC and non-LDCs. These outliers are primarily caused by countries with insufficient data, where Instagram did not consistently rank in the top 50 apps every month. This lack of consistent ranking can cause disproportionately high weighted mean ranks for these countries. Despite these outliers, most of the data is clustered at lower weighted ranks.

In the second box-plot (right), after cleaning, the distributions for LDCs and non-LDCs become more representative. The median and mean values for both groups are relatively close, however non-LDC countries are slightly higher than LDCs. Suggesting higher popularity in non-LDC countries. The median of both groups is smaller than their corresponded mean and the whisk length to the minimum is shorter than that to maximum, which means that the distribution of popularity in both groups is positive skewed.

Overall, after accounting for outliers, the analysis shows some differences in the central tendencies of Instagrams popularity between the two groups, suggesting that Instagram is possibly more popular in non-LDCs.

```
LDC Group Statistics:
count      20.000000
mean       35.591449
std        79.790747
min         3.741935
25%        11.217742
50%        15.516129
75%        24.354839
max       372.000000
Name: weighted_mean_rank, dtype: float64

Non-LDC Group Statistics:
count     120.000000
mean       17.382335
std        44.945588
min         1.225806
25%         7.112903
50%        10.516129
75%        17.766129
max       496.000000
Name: weighted_mean_rank, dtype: float64
```

(II) (b) From the image on the right (using non-cleansed data), on average, Instagrams ranks lower in popularity in LDCs, with a higher mean weighted rank of 35.59 compared to 17.38 for non-LDCs. This trend is further supported by the median values, where the LDC group has a median rank of 15.52 compared to 10.52 for non-LDCs. Additionally, the LDC group shows a greater variability, with a STD of 79.79 compared to

44.95 in the non-LDC group. The LDC group also has a slightly higher minimum rank and a broader overall range of values, suggesting more inconsistency in Instagrams popularity across these countries. With this descriptive statistics, we can still guess that Instagram is still more popular in non-LDCs than in LDCs.

(III) (a) From the visualization and descriptive statistics, Instagram tends to be less popular in LDCs based on the higher average (35.59), median (15.52) and weighted mean ranks for LDCs, compared to non-LDCs (mean: 17.38, median: 10.52). These higher ranks indicate lower popularity, as a higher rank means a lower position in the popularity ranking. This suggests that, on average, Instagram ranks lower in LDCs.

(III) (b) To determine if the observed difference in Instagram popularity between LDCs and non-LDCs is statistically significant, the Levene's test followed by an independent samples t-test is an appropriate approach. This will be calculated using the cleansed dataset.

First the Levene's test is used to see the equality of variances between the two groups. The assumption of equal variances is important for performing the t-test. Levene's test is chosen because it is robust to deviations from normality and tests the null hypothesis that the variances are equal across the two groups.

After confirming the assumptions of equal variance, the independent samples t-test is used to compare the mean weighted ranks between the two groups. This test is appropriate because we are comparing the means of two independent groups, assuming the data is approximately normally distributed.

(III) (c) H0: No significant difference in mean weighted ranks of Instagrams popularity between LDC and non-LDCs.

H1: Significant difference in mean weighted ranks of Instagrams popularity between LDC and non-LDCs.

```
Levene's Test: Variances are not significantly different (p >= 0.05). Using Student's t-test.
```

```
Independent Samples T-Test:
T-statistic: 2.018822513908348, P-value: 0.04553123453968442

The p-value is less than 0.05, indicating a statistically significant difference between the mean weighted ranks of Instagrams's popularity in LDCs and non-LDCs.
```

The p-value is less than 0.05, therefore at 95% confidence level we can reject H0.

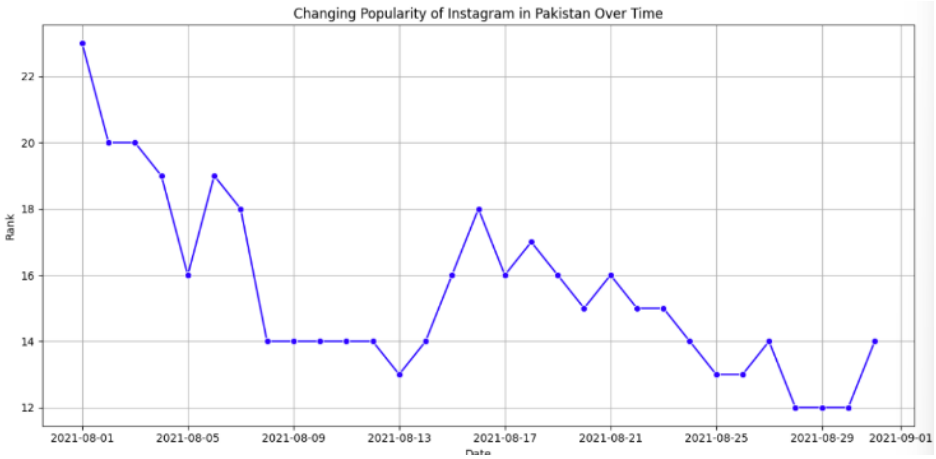
(III) (d) The effect size is measured using Cohen's d (0.51). This represents a medium effect size. This means that while there is a statistically significant difference between the two groups, the magnitude of this difference is moderate in practical terms. A medium effect size shows that LDC status has a noticeable influence on Instagram's app store ranking, with countries classified as LDCs generally having less favorable ranks. However, the effect is not strong enough to suggest a large or substantial gap, meaning other factors beyond LDC status might also contribute to Instagrams popularity differences. This medium effect size highlights a meaningful but not overwhelming disparity between the groups.

```
Cohen's d: 0.5114289302791456
Effect Size: Medium
T-statistic: 2.018822513908348, P-value: 0.04553123453968442
```

(III) (e) The analysis shows that Instagram is less popular in LDCs compared to non-LDCs, as shown by a higher mean weighted rank for LDCs. This difference is statistically significant (T-statistic = 2.02, P-value = 0.046) with a medium effect size (Cohen's d = 0.51). This suggests that LDC status moderately impacts Instagram's popularity. However, the analysis is limited by the scope of the dataset, which may not account for all relevant countries or variations in ranking methodologies. Limitations such as incomplete data, for example: not including every single country in the globe, and other external socio-economic factors like

internet accessibility mean that the results should be interpreted cautiously. The findings point to a noticeable but not dominant disparity, emphasizing that while LDC status is a factor, it is not the sole factor of Instagram's popularity. The results that we have gathered assume that no other factors are significantly related to the ranking in the dataset.

2. (I) (a) The line graph on the right shows the changing popularity of Instagram in Pakistan over the course of August 2021. The x-axis represents the dates, while the y-axis shows Instagram's rank, where a lower rank means higher popularity. At the beginning of the month, Instagram's rank was quite high, peaking at 23 on August 1st, which shows the lowest popularity. During the two weeks in August, the rank fluctuated significantly, indicating a period of instability in Instagram's performance. However, from August 16th onwards, the rank began to show noticeable improvement, with a steady decline in values that show an increase in popularity.



Around mid-August, the graph shows a period of stable decrease, with ranks consistently hovering between 13 and 15. Although there were minor fluctuations during this phase, the general pattern shows that Instagram maintained a higher level of popularity compared to the start of the month. This stable decrease suggests that Instagram's usage in Pakistan was increased through the latter half of the month. By the final days of August, the rank reached its lowest value of 12, showing the platforms peak popularity during the observed period.

Overall, the graph effectively highlights two key phases: an initial period of variability with lower popularity and a later period of stabilization where Instagram's rank consistently improved. The overall downward trend in rank shows that Instagram's popularity in Pakistan increased steadily throughout the month, ending in a period of relative consistency and high popularity.

(I) (b) As seen in the descriptive statistics image on the right, The popularity of Instagram in Pakistan throughout August 2021 exhibited notable fluctuations. The mean rank of 15.48 and the median rank of 15 suggests that Instagram's popularity remained fairly consistent, with ranks clustering closely around these central values, but the data reveals variability. For example, Instagrams rank peaked at 23 early in the month before dramatically improving to a low of 12 by August 10th. However, after mid-August, it fluctuated between ranks 12 and 16 showing a mix of improvements and declines.

	rank
count	31.000000
mean	15.483871
std	2.669219
min	12.000000
25%	14.000000
50%	15.000000
75%	16.500000
max	23.000000

dtype: float64

The IQR from 14 (25th percentile) to 16.5 (75th percentile) suggests that while there were notable fluctuations, the majority of the ranks were concentrated within a relatively tight range in the latter half of the month. The standard deviation of 2.67 reinforces this idea of there being ups and downs, whilst variability in rank was not extreme.

Therefore, while Instagram did not maintain a perfectly steady popularity throughout August, the platform remained fairly popular, with fluctuations showing periods of increased interest as well as some dips in ranking.

(II) (a) Instagram seems to have risen in popularity in Pakistan over this period. This conclusion is evident from the overall trend in the rank data. At the start of the observed period, Instagram had its lowest popularity, with a rank of 23. However, as the month progressed, the rank steadily improved, reaching vales between 12 and 15 during the latter half of the month. This downward trend in rank (where lower ranks mean higher popularity) highlights an increase in Instagrams overall popularity. Overall showing an improved popularity in Pakistan during the month of August 2021.

(II) (b) Spearman's rank correlation test is an appropriate inferential statistical test to determine whether there is a statistically significant difference between time and Instagram's popularity. This non-parametric test is ideal because it does not assume normality in the data, which is particularly useful when the relationship between variables may not be linear. This test is suitable because it evaluates the strength and direction of the relationship between two ranked variables without assuming any specific distribution for the data. Given that Instagram's rank is ordinal and time is continuous, Spearman's test is an ideal choice compared to other tests which may assume linearity.

Overall, Spearman's rank correlation is a robust method to assess the relationship between time and Instagram's rank, providing a statistically valid way to determine whether changes in Instagram popularity are associated with changes in time. The p-value from this test will show whether the observed relationship is statistically significant or whether it could have occurred by chance.

(II) (c) H0: No monotonic relationship between time and popularity for Instagram in Pakistan during August.

H1: There is a monotonic relationship between time and popularity for Instagram in Pakistan during August.

```
P-value: 0.0000338823
The p-value is less than 0.05, indicating a statistically significant monotonic relationship.
```

The p-value is much less than 0.05, so the H0 is rejected at a 95% confidence level. This tells us that the rank of Instagram in Pakistan does decrease as time goes on, meaning it is becoming more popular.

(II) (d) The effect size is measured using Spearman's rank correlation coefficient, calculated at Rho being -0.6727. This reveals a moderately strong negative and monotonic relationship between time and Instagram's popularity in Pakistan during the observed period. This negative value suggests that as time progresses, Instagram's rank decreases, which implies an increase in its popularity. A Rho of -0.6727 is considered a strong effect, showing that the decrease over time is not just coincidental, but part of a consistent trend. The strength of the negative correlation suggests that there is a notable inverse relationship between time and Instagram's rank, supporting the idea that Instagram became more popular as its rank decreased. Overall, the effect size proves a moderately strong relationship between time and Instagram's rising popularity during the period.

```
Spearman's Rank Correlation Coefficient (rho): -0.6727
The negative rho suggests that as time progresses, Instagram's rank decreases (increasing popularity).
```