Assignment 1

1. Choose a real-world dataset from a domain of your interest. Describe the dataset's attributes, size, and source. Perform basic summary statistics on the dataset, and identify any potential challenges.

Ans.

We are choosing the sample dataset provided in the assignment description for our assignment. That dataset is titled - “Online News Popularity.”
The dataset performs the analysis to check a news article’s popularity by considering the number of times an article is shared. The number of shares is correlated by various attributes like the number of HREFs, the number of keywords, the day of the week the article was published, the section in which the article was published (sports, technology, etc.), and others. For detailed information, we have provided the list below, which contain all the attributes in the dataset with their description.

0. URL: URL of the article
1. timedelta: Days between the article publication and the dataset acquisition
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)

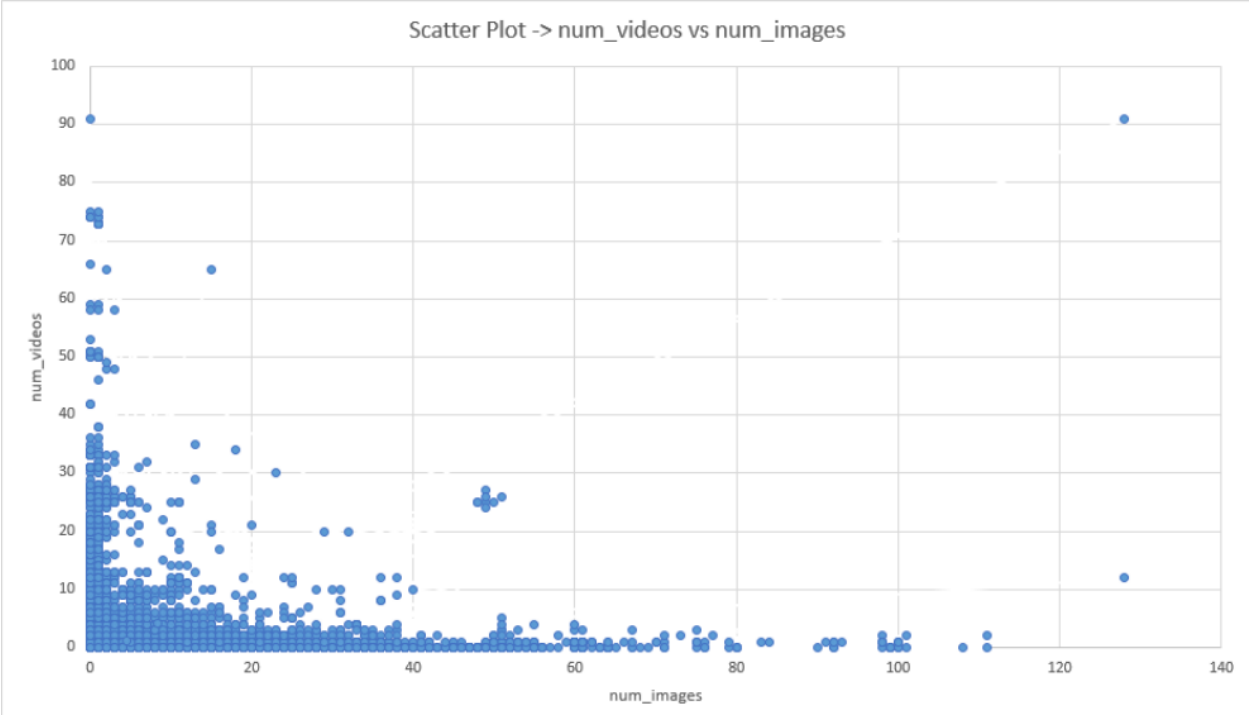
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

Size of the database: The dataset has 39644 data objects and 61 attributes.
Source of the database: The dataset has been picked from the website link.
Summary statistics: The statistics of the dataset have been calculated and collected in the Python and Excel files provided.

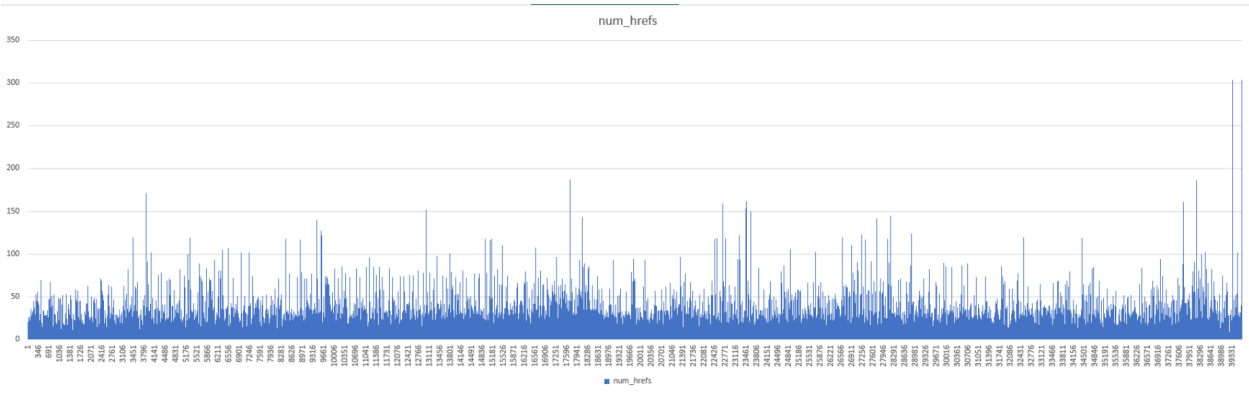
2. Create histograms, box plots, and scatter plots to visualize the distribution of numerical attributes and visually validate the observed patterns.

Ans

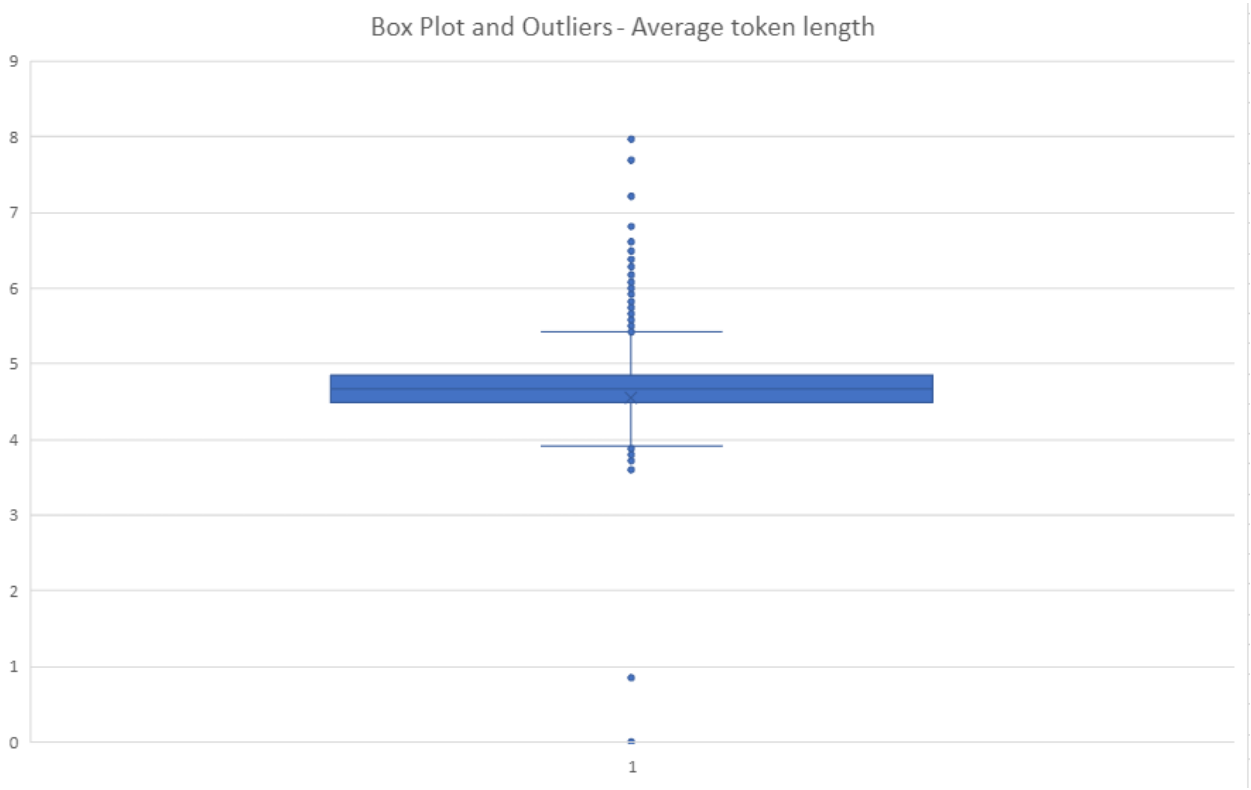
The code for the scatter plots, box plots, and histograms has been provided in the Python file.
Below is a scatter plot made in Excel.



Below is a histogram made in Excel.

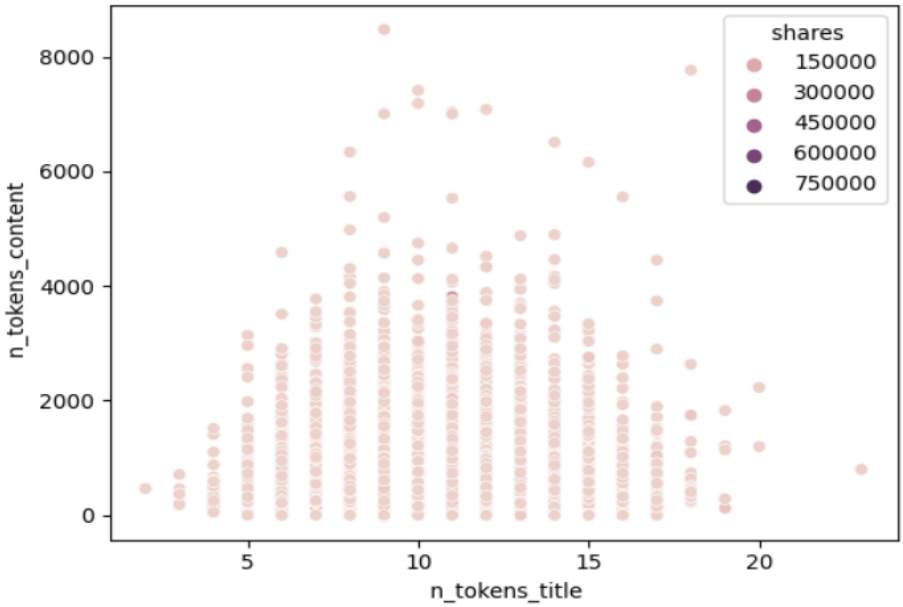


Below is a box plot and outliers made in Excel.

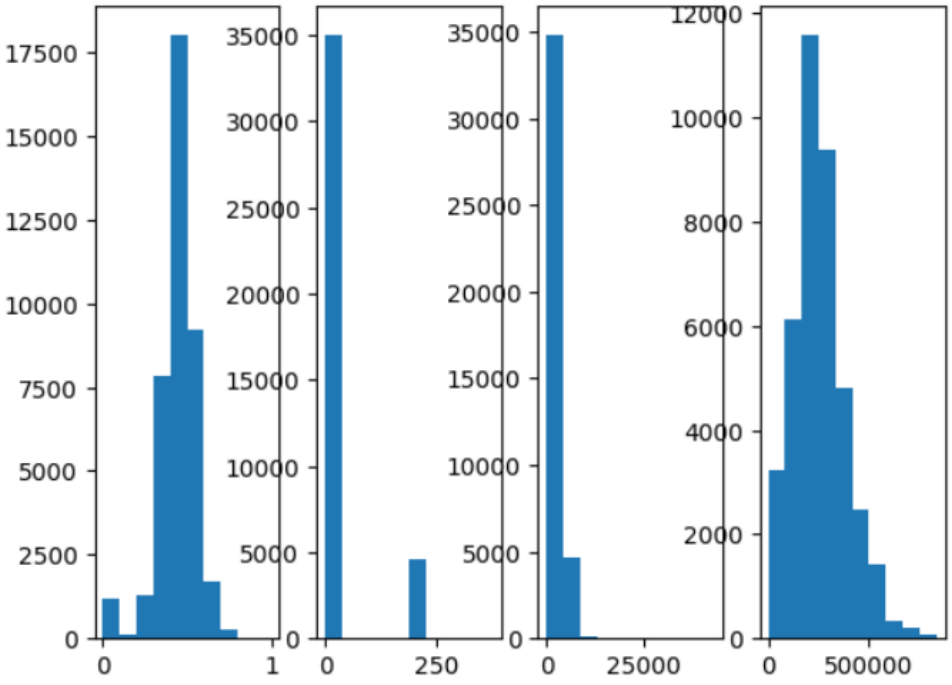


Data visualization using Python:

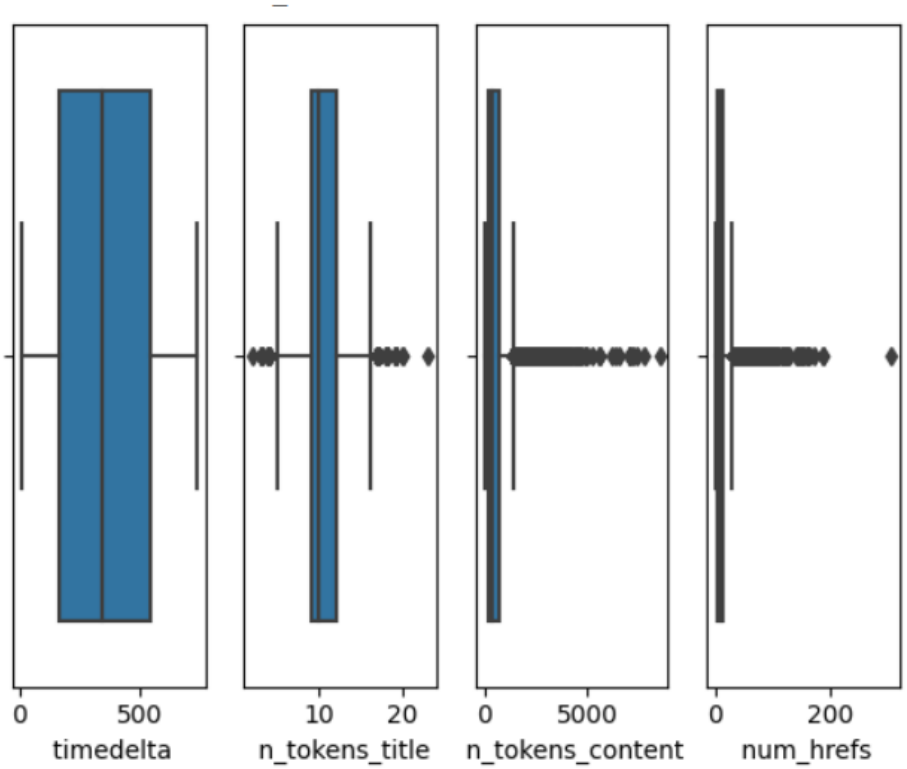
Scatter plot:



Histogram



Box plot



---

3. Explore relationships between pairs of attributes by computing the correlation matrix for this dataset using the formula for the cosine between centered attribute vectors. Output which attribute pairs are i) the most correlated, ii) the most anti-correlated, and iii) the least correlated.

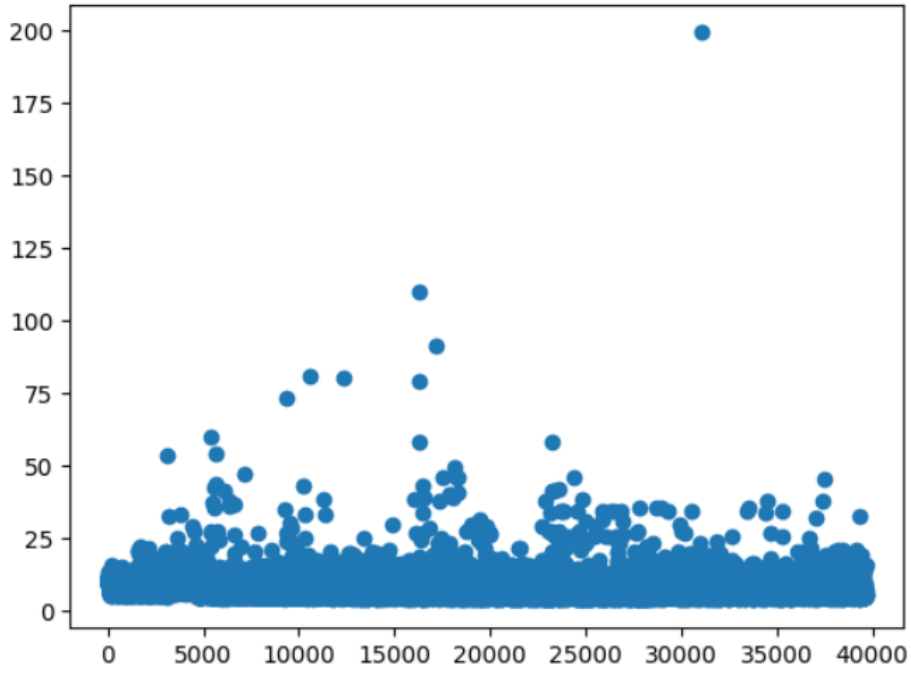
Ans

The correlation was calculated using the inbuilt correlation function provided by the pandas library. The pairwise correlation values are then sorted and used to identify the most correlated pair of attributes(pair having max correlation), least correlated(pair having smallest correlation value which is greater than 0), and most anti-correlated(pair having the least correlation value, in this case, the correlation would be negative).
To calculate correlation manually, data needs to be centered. The Dot product of this centered data with its transpose divided by the square of the norm of the matrix gives the correlation matrix.

4. Investigate any outliers or anomalies and discuss their potential impact on Analysis.

Ans

Anomalies were found by calculating the Mahalanobis distance, and the points with significantly different Mahalanobis distance were considered outliers. (as visible in the plot).



If the threshold for outliers is set as 100, then there are two outliers.

Outliers can impact the analysis quite severely. Since they have extreme values, they can substantially change summary stats like mean, range, etc. Outliers can also lead to some problems in data visualization, as the scale might need to be adjusted to accommodate the outliers. Also, outliers can cause classification models to have greater errors.

5. Compute the mean vector μ for the data matrix, and then compute the total variance.

Ans.

The mean vector μ for the data matrix can be found in Excel by finding the mean for each attribute. This vector μ can be seen in the dataset at the bottom.
Similarly, the variance can be calculated in Excel, which is shown below the mean vector μ in the Excel sheet.

6. Calculate the sample covariance matrix, denoted as Σ, by evaluating the inner products among the attributes of the data matrix that has been centered. Subsequently, determine the sample covariance matrix by summing the outer products of the centered data points.

Ans.

This has been done inside the Python file itself so that the results can be found there.

For the Excel, this has been done in the excel sheet.

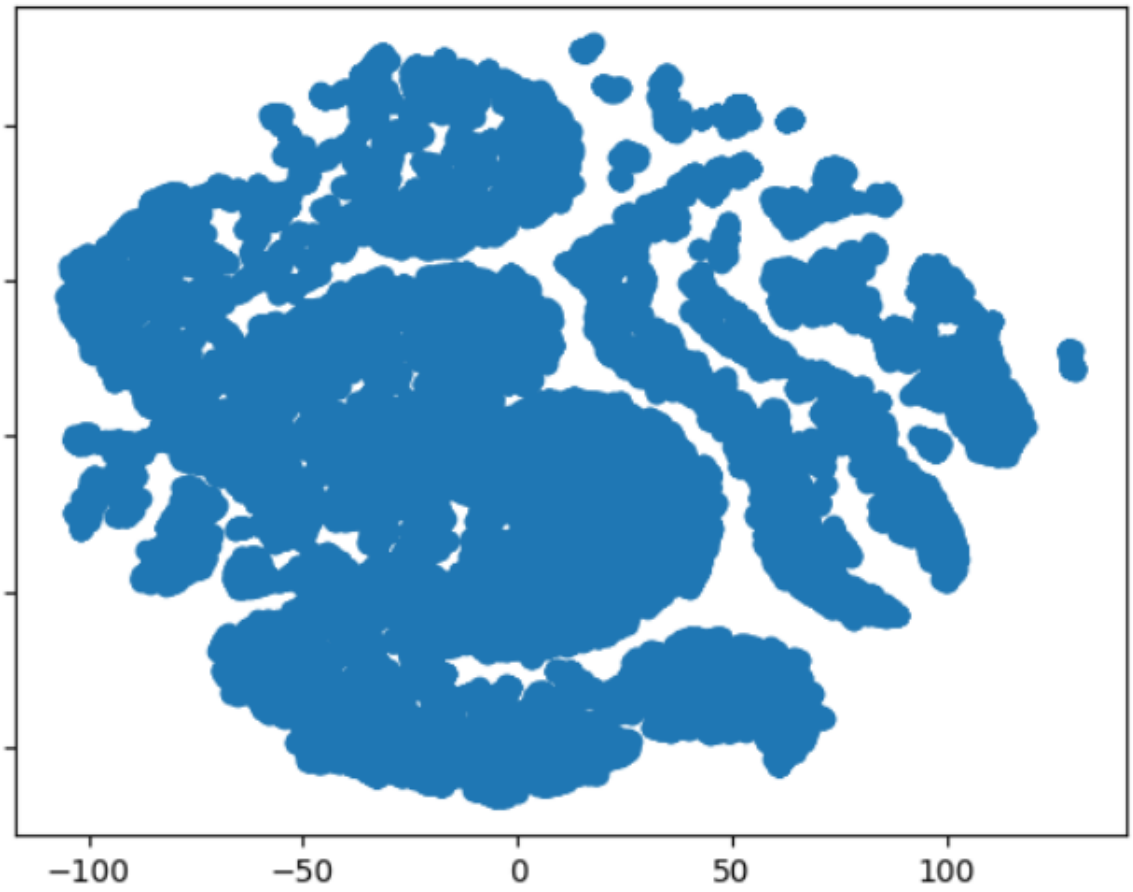
Since our database has more than 39000 data points and 61 attributes, Excel was unable to make the covariance matrix of the whole dataset. So we have made the matrix for the first 10000 data points only for all the 61 attributes. This can be seen in the Sheet 2 (named as Covariance) in the excel file.

Dimensionality Reduction:

1. Apply dimensionality reduction techniques like Principal Component Analysis (PCA) or t-SNE to visualize high-dimensional data in lower dimensions.

Ans.

TSNE with two components: Made scatter plot



2. Interpret the results of the dimensionality reduction and discuss whether it helped reveal any patterns or clusters.

Ans.

As you can see, 2 clusters have been formed. These two attributes account for a large amount of variance in the data.