**Part I: Research Question**

A.  Describe the purpose of this data mining report by doing the following:

1.  Propose **one** question relevant to a real-world organisational situation that you will answer using principal component analysis (PCA).

A. I will use the PCA to investigate relevant features in identifying customers at high risk of churn.

2.  Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scenario's scope and represented in the available data.

A. The stakeholders can review the data provided by the analysis and create incentives to keep the customers likely to terminate their contracts with the company. This will lead to a lower churn rate.

**Part II: Method Justification**

B.  Explain the reasons for using PCA by doing the following:

1.  Explain how PCA analyses the selected data set. Include expected outcomes.

A. PCA computes transformed variables along the direction of significant variability in the data set. Fewer dimensions can account for most of the variability using the new dimensions, designed to be along the direction of variability. The predictive analytics or classification analysis can run on fewer dimensions without losing essential details. Chantal D. Larose, & Daniel T. Larose. (2019).

Expected outcomes: I expect that the PCA will provide a list of features from the original dataset relevant to identifying customers at high risk of churn.

2.  Summarize **one** assumption of PCA.

A. The data set variables must correlate for PCA to identify them for dimension reduction.

**Part III: Data Preparation**

C.  Perform data preparation for the chosen dataset by doing the following:

1.  Identify the continuous dataset variables that you will need to answer the PCA question proposed in part A1.

1. 'Outage\_sec\_perweek', 'Tenure', 'MonthlyCharge', 'Bandwidth\_GB\_Year'

2.  Standardize the continuous dataset variables identified in part C1. Include a copy of the cleaned dataset.

A. Graphical user interface, text, application

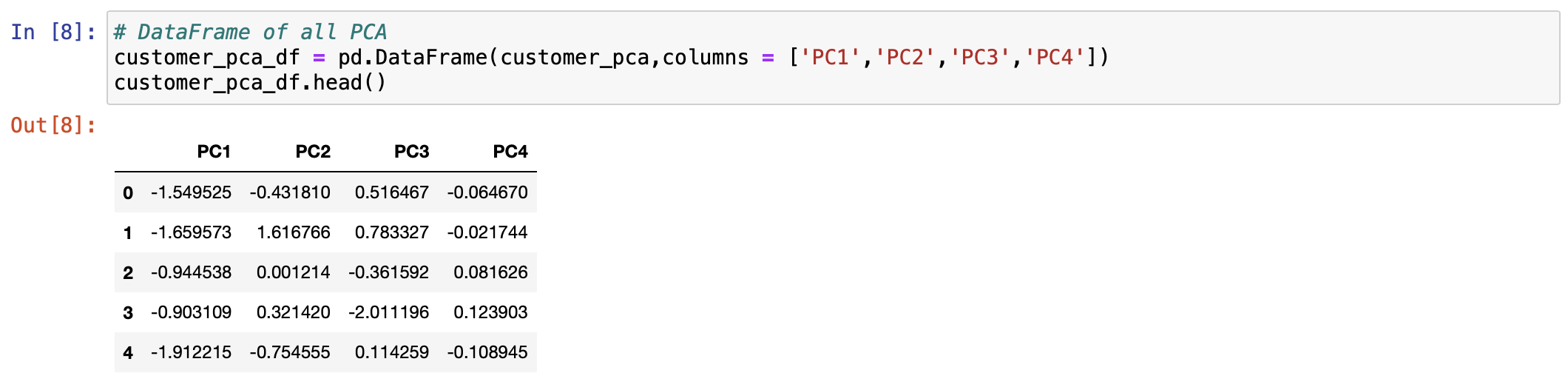
Description automatically generated

B. Attached is‘ prepared\_dataset.csv’

**Part IV: Analysis**

D.  Perform PCA by doing the following:

1.  Determine the matrix of all the principal components.

A. 

2.  Identify the total number of principal components using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.

A.

Chart, line chart

Description automatically generated

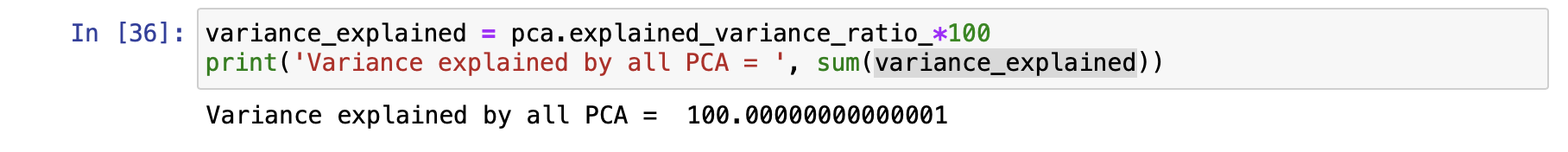
3.  Identify the variance of each principal component in part D2.

A.

Text

Description automatically generated

4.  Identify the total variance captured by the principal components in part D2.

A. 

5.  Summarize the results of your data analysis.

A.

Based on the analysis, the predictors mentioned in the C2 can explain 100% of the variability. This can be seen in the scree plot as it shows that two predictors can explain most of the variance, with the elbow plot sharply going to 100 at two. Lastly, all four predictors show 100% of the variance when summed up. This is insignificance analysis, and I recommend we need additional data to make sense and provide predictors that can identify customers that are about to churn.

**Part V: Attachments**

E. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

https://developers.google.com/. (2019, November 29). *Clustering Algorithms | Clustering in Machine Learning |*. Google Developers. Retrieved June 18, 2022, from https://developers.google.com/machine-learning/clustering/clustering-algorithms

GeeksforGeeks. (2018, November 28). *How to select multiple columns in a pandas dataframe*. https://www.geeksforgeeks.org/how-to-select-multiple-columns-in-a-pandas-dataframe/

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists : 50+ essential concepts using r and python*. O'Reilly Media, Incorporated.

Chantal D. Larose, & Daniel T. Larose. (2019). *Data Science Using Python and R*. Wiley.