Part I: Research Question

*A. Describe **one** question or decision that you will address using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.*

- Is there a correlation between extra service purchases and churn?

B. Describe the variables in the data set and indicate the specific type of data being described. Use examples from the data set that support your claims.

The data types in this set include:

- Int64: 64-bit integers
- Float64: 64-bit integer with a decimal point
- Object: can contain multiple types of data

caseorder	int64
customer_id	object
interaction	object
city	object
state	object
county	object
zip	int64
lat	float64
lng	float64
population	int64
area	object
timezone	object
job	object
children	float64
age	float64
education	object
employment	object
income	float64
marital	object
gender	object
churn	object
outage_sec_perweek	float64
email	int64
contacts	int64
yearly_equip_failure	int64
techie	object
contract	object
port_modem	object
tablet	object
internetservice	object
phone	object
multiple	object
onlinesecurity	object
onlinebackup	object
deviceprotection	object
techsupport	object

```
streamingtv            object
streamingmovies         object
paperlessbilling        object
paymentmethod           object
tenure                 float64
monthlycharge           float64
bandwidth_gb_year       float64
item1                    int64
item2                    int64
item3                    int64
item4                    int64
item5                    int64
item6                    int64
item7                    int64
item8                    int64
index                    int64
dtype: object
```

Part II: Data-Cleaning Plan

C. Explain the plan for cleaning the data by doing the following:

1. Propose a plan that includes the relevant techniques and specific steps needed to identify anomalies in the data set

   - Change all characters in the CSV file headers to lowercase.
   - Make sure all character cases and field types are consistent within the CSV columns.
   - Remove unnecessary spaces, random characters from the data set.
   - Import pandas, numpy, scipy, and the churn dataset.
   - Create and assign an index.
   - Re-express categorical data as numeric.
   - Standardize the numeric fields.
   - Identify outliers using z-scores.
   - Identify null values

2. Justify your approach for assessing the quality of the data, include the characteristics of the data being assessed, and the approach used to assess the quality:

   - Completeness: how comprehensive is my data?

     Look at the variables and judge if all important information is there.

   - Accuracy: does the information mirror a real-life scenario?

     Identify and count null values, determine z-scores and quantify outliers.

3. Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

   I will be using python as my selected programing language. Libraries include:

   - Pandas to import the file.
   - Numpy to identify and impute null values.
   - Scipy to calculate z-scores.

4. Provide the code you will use to identify the anomalies in the data.

   See D206 Keim Task One Code document

Part III: Data Cleaning

D. Summarize the data-cleaning process by doing the following:

1. Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.

   Nulls were found in the following columns: children, age, income, techie, phone, techsupport, tenure, bandwidth_gb_year.

   Outliers based on z-scores were found in the following columns: population, outage_sec_perweek, email, contacts, yearly_equip_failure, and monthlycharge. Each column's outliers accounted for less than 5% of the dataset.

2. Justify your methods for mitigating each type of discovered anomaly in the data set.

   I first dropped outliers. Because each of the z-score column's outliers accounted for less than 5% of the dataset, I dropped all of the outliers identified.

   To impute missing data, I first dropped any columns containing all null values. I then considered which columns were most important to the dataset and decided that the columns indicating if a customer has extra services were the most important columns that contained nulls. I dropped any rows that had nulls in all the extra services columns. Then, I forward filled the remaining nulls. Then I backfilled the last few.

5. Summarize the outcome from the implementation of *each* data-cleaning step.

   - Change all characters in the CSV file headers to lowercase.

     All characters in the header were changed to lowercase to create a tidy table before importing. This allowed for easier coding in python.

   - Make sure all character cases and field types are consistent within the CSV columns.

     All alphabetical characters in the dataset were changed to consistent cases within their columns to create a tidy table before importing. All numeric value was changed to a number type. This allowed for easier coding in python.

   - Remove unnecessary spaces, random characters from the data set.

This created a tidy table before importing, allowing for easier coding in python.

- Import pandas, numpy, scipy, and the churn dataset.

  Importing these libraries allowed their functions to be used in python. Importing the dataset made it available to be manipulated in python.

- Create and assign an index.

  This created a new numeric variable numbering the rows from 0-9999 to allow easier organization of the dataset.

- Re-express categorical data as numeric.

  Selected variables with an object datatype were converted to a numeric datatype. It is easier to examine quantified data in these variables than if they are left as strings.

- Standardize the numeric fields.

  Selected numeric fields were calculated into z-scores to better identify outliers

- Identify outliers using z-scores.

  Using z-scores, I returned queries of the calculated numeric fields to identify outliers. The threshold used was any datapoint with a z-score absolute value greater than 3.

- Identify null values.

  I returned the count of nulls in each column, and the total count for the dataset.

6. Provide the code used to mitigate anomalies.

   See D206 Keim Task One Code document

7. Provide a copy of the cleaned data set.

   See D206 Keim Task One Clean Churn Data document
   .

7. Summarize the limitations of the data-cleaning process.

   When cleaning survey data, like this dataset, you must impute your best guess in place of missing or misleading data. You cannot guess a truly accurate answer from the survey subject. Human error is a large factor in missing data in these sets and there is no way to have 100% accuracy in addressing those values.

   Cleaning data with an object datatype is also difficult as not all of it can be quantified, especially when the survey asked for a ‘write in’ answer. In cases such as these, it doesn’t do much good to re-express the categorical data as numeric because it is too varied and cannot be accurately standardized. In these cases, we ask ourselves if the variable is genuinely relevant to the business question.

8. Discuss how the limitations in part D6 affect the analysis of the question or decision from part A.

   As discussed above, several variables cannot be easily quantified. Address data, GPS data, and education level cannot be easily standardized using the methods above. These factors may have some influence on the relationship between the purchase of extra services and churn, but quantifying that effect is limited.

E. Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:

1. List the principal components in the data set.

   The principal components are population, children, age, income, outage_sec_perweek, contacts, yearly_equip_failure, and tenure.

2. Describe how you identified the principal components of the data set.

   I created a scree plot of the eigenvalues and identified the components with eigenvalues of 1 or higher.

3. Describe how the organization can benefit from the results of the PCA

   The results of the PCA can further reduce data after it has been cleaned. Variables can be grouped statistically, and the PCA identifies the most impactful ways to do that.

Bibliography

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

Ekker, R.-J. (n.d.). *Pandas Playbook: Manipulating Data.* Retrieved from https://app.pluralsight.com/player?course=pandas-playbook-manipulating-data&author=reindertjan-ekker&name=243c075e-6df5-4667-8293-d4b842c76774&clip=0&mode=live

Vallisneri, M. (n.d.). *Python Statistics Essential Training:.* Retrieved from https://www.linkedin.com/learning/python-statistics-essential-training/welcome?u=2045532

Western Governor's University. (n.d.). *Principal Component Analysis and Data Reduction*. Retrieved from https://wgu.ucertify.com/?func=ebook&chapter_no=6#top