The Forage is a website where companies' recruitment teams create virtual work experiences/job simulations to educate users and find potential candidates. Users can gain practical work experience to build up their skills and showcase their abilities. I discovered this platform through Twitter and thought it would be an excellent opportunity to practice and demonstrate my skills as I transition into a data analyst role. I decided to start with the Accenture Data Analytics and Visualization virtual experience program.

The project encompassed four main tasks: understanding the project, cleaning and modeling the data, visualizing the data, creating a narrative, and presenting the data visualizations and story to the client. Throughout the program, you are provided with videos and documents to offer context and guidance for completing the tasks and specifics about your role as a data analyst.

To kickstart the understanding of the projects and the goals of the company you're working for, there is a video from an Accenture employee explaining the objectives of the project task. You become aware of the roles and responsibilities of a data analyst and the impact of your work on the business. Similar to a real-world scenario, you are given a brief from the company you're working with, known as Social Buzz. Your task is to read the brief and take note of crucial information because there is a quiz. In the real world, this information is essential for finding solutions for the client. You are also provided with an organizational chart, which simulates a team meeting. This introduction allows you to become familiar with the names and roles of your team members, akin to a kickoff meeting at the start of a new project.

Once I understood my role in the project, it was time to work with the actual data. There were seven data sets displayed via a data model that visually explained the relationships between the datasets. The data model was a helpful tool for illustrating how data could be merged to create new and more useful datasets. Three data files were provided as CSV files. Before merging all three datasets, each dataset had to be prepared and processed. I opted to use a Google Colab notebook and Python for my data cleaning process. While the types of cleaning steps were suggested, it was up to me as the data analyst to determine which datasets required specific cleaning processes.

Data Cleaning Process

- Cleaning the Content table
    - dropping/removing columns that didnt contain useful/relevant data also checked for missing data

Drop Irrelavant Columns: `Unnamed: 0`, `URL`, User ID`

This columns don't provide the kind of insights that we're looking to share with and find for the client therefore they can be removed.

```
[ ] content = content.drop(columns=['Unnamed: 0', 'URL','User ID'])
```

```
[ ] content.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Content ID  1000 non-null   object
 1   Type        1000 non-null   object
 2   Category    1000 non-null   object
dtypes: object(3)
memory usage: 23.6+ KB
```

-
    - Checked the Categories column and found there were many duplicates because of various of spelling and capitalizations

```python
content['Category'].value_counts()
```

```
technology           71
animals              67
travel               67
culture              63
science              63
fitness              61
food                 61
healthy eating       61
cooking              60
soccer               58
tennis               58
education            57
dogs                 56
studying             55
veganism             48
public speaking      48
Fitness               5
Animals               4
Science               4
"soccer"              3
"culture"             3
Soccer                3
"dogs"                2
Education             2
Studying              2
Travel                2
Food                  2
"veganism"            1
"public speaking"     1
Public Speaking       1
"technology"          1
"cooking"             1
Healthy Eating        1
"studying"            1
"food"                1
Culture               1
"tennis"              1
Technology            1
"animals"             1
Veganism              1
"science"             1
Name: Category, dtype: int64
```

- Removed quotations and make all the terms lowercased.

```python
content['Category'] = content['Category'].str.replace('"','')
content['Category'] = content['Category'].str.lower()
```

- Verified that the Categories were uniform and counted correctly.

```python
content['Category'].value_counts()
```

```
technology         73
animals            72
travel             69
science            68
culture            67
fitness            66
food               64
soccer             64
healthy eating     62
cooking            61
tennis             59
education          59
studying           58
dogs               58
public speaking    50
veganism           50
Name: Category, dtype: int64
```

- Changed column name for more specification, each table has a Type column change Type to Content_Type

Change `Type` to `Content_Type`

This makes it clear that there is a difference between `Type` in the content table and `Type` in the reactions or reactions_type tables

```
[ ] content.rename(columns = {'Type':'Content_Type'}, inplace=True)
    content.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Content ID    1000 non-null   object
 1   Content_Type  1000 non-null   object
 2   Category      1000 non-null   object
dtypes: object(3)
memory usage: 23.6+ KB
```

- Cleaning Reactions tables
  - Dropped irrelevant columns
  - Checked for missing data and remove rows with missing

```
] reactions = reactions.dropna(subset=['Type'])
```

  - Changed data type for Datetime column

Change `Datetime` to datetime Dtype

Changing the `Datetime` column to the datetime data type allows us to use functionalities specific to datetime

```
[ ] reactions['Datetime'] = pd.to_datetime(reactions['Datetime'])
    reactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 24573 entries, 1 to 25552
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Content ID  24573 non-null  object
 1   Type        24573 non-null  object
 2   Datetime    24573 non-null  datetime64[ns]
dtypes: datetime64[ns](1), object(2)
memory usage: 767.9+ KB
```

  - Changed Type column to Reaction_Type for clarity
- Cleaning reaction_types table
  - Dropped irrelevant columns
  - Changed Type column to Reaction_Types for clarity
- Merging all the tables into one table
  - Merged reactions table to content => reactions_content_merge

```
reactions_content_merge = reactions.merge(content,how='inner', on=['Content ID'])
reactions_content_merge.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 24573 entries, 0 to 24572
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Content ID     24573 non-null  object
 1   Reaction_Type  24573 non-null  object
 2   Datetime       24573 non-null  datetime64[ns]
 3   Content_Type   24573 non-null  object
 4   Category       24573 non-null  object
dtypes: datetime64[ns](1), object(4)
memory usage: 1.1+ MB
```

  - Merged reactions_content_merge with reaction_types tables

```
merge_all = reactions_content_merge.merge(reaction_types, how='inner', on='Reaction_Type')
```

```
merge_all.shape
```

```
(24573, 7)
```

```
merge_all.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 24573 entries, 0 to 24572
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Content ID     24573 non-null  object
 1   Reaction_Type  24573 non-null  object
 2   Datetime       24573 non-null  datetime64[ns]
 3   Content_Type   24573 non-null  object
 4   Category       24573 non-null  object
 5   Sentiment      24573 non-null  object
 6   Score          24573 non-null  int64
dtypes: datetime64[ns](1), int64(1), object(5)
memory usage: 1.5+ MB
```

After cleaning and merging the data into a single table, I began exploring the data to derive insights. I formulated questions that related to the client's brief. The project brief indicated that the company was preparing for an IPO and wanted an analysis of their top five content categories with the largest total popularity. I evaluated the content categories with the most positive and negative sentiments and identified the most popular types of content. I later used these insights in my presentation.

Following the data exploration, insight discovery, and data visualization, it was time to create a presentation to communicate the findings with the client in a concise and understandable manner. I used Canva to design and record my presentation, which provided valuable practice for presenting to an actual audience.

In conclusion, my virtual internship experience was immensely valuable. It offered a hands-on opportunity to execute the data analysis process while receiving tips and hints along the way. It also served as an excellent platform to showcase my skills in data cleaning, modeling, visualization, and exploration, as well as communication and data storytelling.