### Data Science Blog Post Q&A Template

#### 1. Introduction / Hook

- What inspired you to start this project? Was it a personal curiosity, a real-world issue, or a class assignment?  
Ans: **Edulytics** is a forward-thinking ed-tech analytics startup aiming to revolutionize how educational institutions understand and support their student populations

- Why do you find this topic interesting or important within data science or your field of study?

- How could this project potentially impact others or the industry?

#### 2. Problem Statement

- What specific problem or question did you aim to solve or answer with this project?

Ans: One of their primary goals is to help schools and policymakers make informed decisions by identifying underlying behavioral patterns and interest groups among students.

Design and evaluate an unsupervised clustering model that groups students into distinct clusters based on:

• Online interest indicators (e.g., sports, music, religion, fashion, substance-related terms, etc.)

• Basic demographics (age, gender, grad year, number of friends)

The ultimate goal is to provide a clustering solution that is:

* • **Interpretable**: Understand what defines each group
* • **Actionable**: Inform decisions related to student engagement and support
* • **Valid**: Grounded in real, distinguishable patterns within the data

Understanding clusters of students with similar interests can help:

* • Tailor academic and extracurricular programs
* • Develop targeted counseling and support services
* • Predict potential behavioral trends
* • Foster inclusive and diverse school environments

- Why does this problem matter? Who are the stakeholders or end-users affected?

- How does solving this problem contribute to the broader context of data science or your area of interest?

#### 3. Data

- What dataset(s) did you use? Give details on the source, size, and key variables or features.

Ans: **Dataset Overview**

**Download:** https://www.kaggle.com/datasets/zabihullah18/students-social-network-profile-clustering

**Structure: The dataset contains 41 columns including:**

• **Demographics**: gradyear, gender, age, NumberOffriends

• **Interest Terms** (37 columns): Frequency counts of mentions such as football, shopping, drugs, music, jesus, rock, death, etc.

All features are numerical (either counts or categorical encoded), making the dataset relatively clean and ready for vector-based clustering techniques.

- What preprocessing or cleaning steps were necessary? Were there any notable challenges during this?

Ans:

* There were about 1337 null values in the gender column and 2496 null values in the age column.
* The distribution of the gender is not balanced and for age there are outliers and the age is in decimal
* I filled the missing values in the gender column using :

gender\_cols=['basketball', 'football', 'soccer', 'sports', 'softball', 'volleyball', 'swimming', 'cheerleading', 'mall',

'shopping', 'clothes', 'dress', 'baseball', 'sports', 'drunk', 'hair', 'drugs']

data['gender']=data.groupby(gender\_cols)['gender'].transform(lambda x: x.fillna(x.mode()[0] if not x.mode().empty else 'Unknown'))

* After that I dropped 224 rows with unknown because we are interested in only two genders
* I filled the missing values in the age column and solve the decimal issue using:

data['age']=pd.to\_numeric(data['age'], errors='coerce')

data['age']=data.groupby('gradyear')['age'].transform(lambda x: x.fillna(x.median()))

data['age']=data['age'].round().astype('int')

* The only categorical column is the gender column so I used labelencoder to encode the column
* Because of the outliers in the dataset I used:

#Putting bounds on columns containing outliers

data['age']=data['age'].clip(lower=12, upper=22)

data['sex']=data['sex'].clip(lower=0, upper=22)

data['god']=data['god'].clip(lower=0, upper=39)

data['blonde']=data['blonde'].clip(lower=0, upper=16)

def cap\_outliers(series, lower\_quantile=0.01, upper\_quantile=0.99):

lower\_bound = series.quantile(lower\_quantile)

upper\_bound = series.quantile(upper\_quantile)

return series.clip(lower=lower\_bound, upper=upper\_bound)

data['NumberOffriends']=cap\_outliers(data['NumberOffriends'])

- Did you discover any unexpected or interesting details, anomalies, or missing information in the data?

Ans:

* Top 10 interest terms ranking by frequency:
* bible 306
* marching 624
* abercrombie 753
* drugs 885
* hollister 1035
* tennis 1303
* drunk 1316
* cheerleading 1496
* kissed 1502
* baseball 1510
* dtype: int64
* Bottom 10 interest terms ranking by frequency:
* football 3717
* mall 3736
* basketball 3842
* band 4472
* cute 4791
* shopping 5208
* hair 6232
* dance 6311
* god 6998
* music 10790

#### 4. Approach / Methodology

- Which data science methods, algorithms, or models did you apply? (The inertia for the KMeans Model is:

44526.64534502264

==================================================

The silhoutte score for the KMeans Model is:

0.34532084361419774

==================================================

The silhoutte score for the DBSCAN Model is:

0.7651995252609586

==================================================

The silhoutte score for the AgglomerativeClustering Model is:

0.2411878474210889

- Why did you choose these approaches? Did you try alternative methods?

- How did you prepare your data for modeling (e.g., feature engineering, scaling)?

Ans:

* I scaled the data using StandardScaler:

#Scaling the dataset

scaler=StandardScaler()

columns=data.columns #So as to not lose the column title, we store it

scaled\_data=scaler.fit\_transform(data)

data=pd.DataFrame(data=scaled\_data, columns=columns)

data.head()

- What metrics did you use to evaluate your model’s performance? Why are these metrics appropriate?

Ans:

* Silhouette score and inertia(for kmeans)

#### 5. Results & Insights

- What were the key outcomes or findings? Include quantitative metrics or qualitative patterns.

Ans:

* Top 5 Interest Terms Defining Each Cluster:
* Cluster cluster\_0:
* music 0.558668
* dance 0.385482
* shopping 0.363938
* god 0.334935
* basketball 0.278091
* Name: cluster\_0, dtype: float64
* Cluster cluster\_1:
* hair 1.413333
* music 1.314133
* dance 1.066667
* cute 0.912533
* shopping 0.868800
* Name: cluster\_1, dtype: float64
* Cluster cluster\_2:
* god 5.302752
* jesus 2.362385
* church 2.091743
* music 1.110092
* bible 0.853211
* Name: cluster\_2, dtype: float64
* Cluster cluster\_3:
* hair 4.427984
* sex 3.563786
* music 3.308642
* kissed 1.979424
* rock 1.934156
* Name: cluster\_3, dtype: float64
* Cluster cluster\_4:
* music 0.610209
* god 0.337652
* band 0.304870
* dance 0.244302
* cute 0.181236
* Name: cluster\_4, dtype: float64
* Demographic Mean Values for Each Cluster:
* age gender\_encoded gradyear NumberOffriends
* clusters
* cluster\_0 16.414816 0.126119 2008.331787 33.209480
* cluster\_1 16.937600 0.065600 2007.838933 39.506133
* cluster\_2 17.486239 0.238532 2007.357798 39.490826
* cluster\_3 17.271605 0.074074 2007.567901 29.407407
* cluster\_4 18.315798 0.262566 2006.603185 22.273181
* Gender Counts per Cluster:
* gender F M
* clusters
* cluster\_0 5273 761
* cluster\_1 1752 123
* cluster\_2 166 52
* cluster\_3 225 18
* cluster\_4 4724 1682

- Were there any results that surprised you or challenged your expectations?

- How would you explain your results to someone with little to no data science background?

#### 6. Challenges & Decisions

- What were the biggest challenges faced during the project? (e.g., data quality, model tuning, computational resources)

- How did you approach solving these problems? What trade-offs or decisions did you make?

- If you were to repeat the project, what would you do differently to improve the process or outcomes?

#### 7. Applications

- How can your findings or model be applied practically? (e.g., in business, agriculture, healthcare)

- Who stands to benefit from your work and how?

- Are there any ethical considerations or limitations in using your results?

#### 8. Conclusion & Next Steps

- What did you learn overall from completing this project—technically, conceptually, or personally?

- What future work, additional analyses, or extensions would you like to explore?

- End with an engaging thought or question for your readers to reflect on.

\*\*\*

> Using my answers to the questions above, please write a clear, engaging, storytelling-style blog post suitable for beginners but insightful for intermediate readers. Structure it with sections: Introduction, Problem Statement, Data, Approach, Results, Challenges, Applications, and Conclusion. Ensure smooth transitions, keep the tone friendly and informative, and make the narrative flow naturally.