CS210 PROJECT REPORT

Serra Bayraktar

INTRODUCTION

In an era where the fabric of social interaction is intricately woven with digital threads, the impact of social media on various aspects of life has become an undeniable phenomenon. Among the multifaceted influences of social networking platforms, the realm of fashion stands out, with trends being born and nurtured through likes, shares, and retweets. This project, titled "How Social Media Affects Fashion Trends," aims to elucidate the relationship between the dynamic world of social media and the ever-evolving fashion industry.

The premise of this research revolves around the hypothesis that social media possesses a significant positive influence on fashion trends, shaping the sartorial choices of individuals across the globe. To substantiate this hypothesis, a comprehensive data-driven approach was employed, encompassing a detailed analysis of responses derived from questionnaires distributed among a diverse demographic. The questionnaire aimed to capture the pulse of the public's perception of social media's sway over fashion trends.

An intricate fusion of statistical methodologies and hypothesis testing formed the cornerstone of this study. The analytical journey embarked upon in this project was not just about proving a point but about unraveling the complex tapestry of social media's impact on fashion, thread by thread. The findings of this study are expected to offer insightful revelations about the magnitude and nature of social media's influence on fashion trends, thereby contributing to the broader understanding of digital media's role in shaping contemporary culture.

As we delve into the subsequent sections of this report, we will uncover the methodological framework, data analysis techniques, and the compelling revelations drawn from the hypothesis testing conducted on the questionnaire data. This project is not just an academic endeavor but a narrative about the symbiotic relationship between digital platforms and fashion, a tale of influence and adaptation in a digitally dominated era.

HYPOTHESIS

The core premise of this research pivots on the hypothesis that social media exerts a positive influence on fashion trends and significantly impacts the fashion choices of individuals. This hypothesis is grounded in the observation of the pervasive role that social media platforms play in the dissemination and popularization of fashion-related content. It is postulated that the interactive and highly visual nature of social media catalyzes the spread of fashion trends, thereby shaping the collective fashion consciousness of its users.

Hypothesis Statement:

To articulate this concept more formally, the hypothesis can be stated as follows:
"H1: Social media has a positive influence on fashion trends and significantly sways the fashion preferences and choices of individuals."

This hypothesis operates on the assumption that there is a measurable and positive correlation between exposure to fashion-related content on social media and the adoption of those trends by individuals. It suggests that social media is not merely a reflective medium for existing fashion trends but a formidable force in creating and fostering new fashion movements.

MATERIALS

1. Google forms (for collecting data)
2. Visual Studio

DATA

Data Collection Method

1. **Questionnaire Design**: The questionnaire was crafted with a focus on comprehensiveness and clarity, ensuring that the questions were both exhaustive and easily understandable. It encompassed a range of question types, including multiple-choice, Likert scale, and open-ended responses, to glean nuanced insights into the respondents' attitudes and behaviors concerning fashion trends and social media.
2. **Distribution and Response**: The Google Form was shared across various social media platforms to garner a diverse set of responses. The target demographic was broad, aiming to capture a wide spectrum of perspectives across different age groups, genders, and cultural backgrounds. A total of 50 individuals participated in the survey, providing a rich dataset for analysis.
3. **Data Compilation**: Upon completion of the data collection phase, the responses were automatically compiled into a CSV (Comma-Separated Values) file by Google Forms. This format facilitated the seamless import and manipulation of the data in various data analysis software and programming environments.

Data Description

The dataset derived from the questionnaire is comprehensive and structured, consisting of the following components:

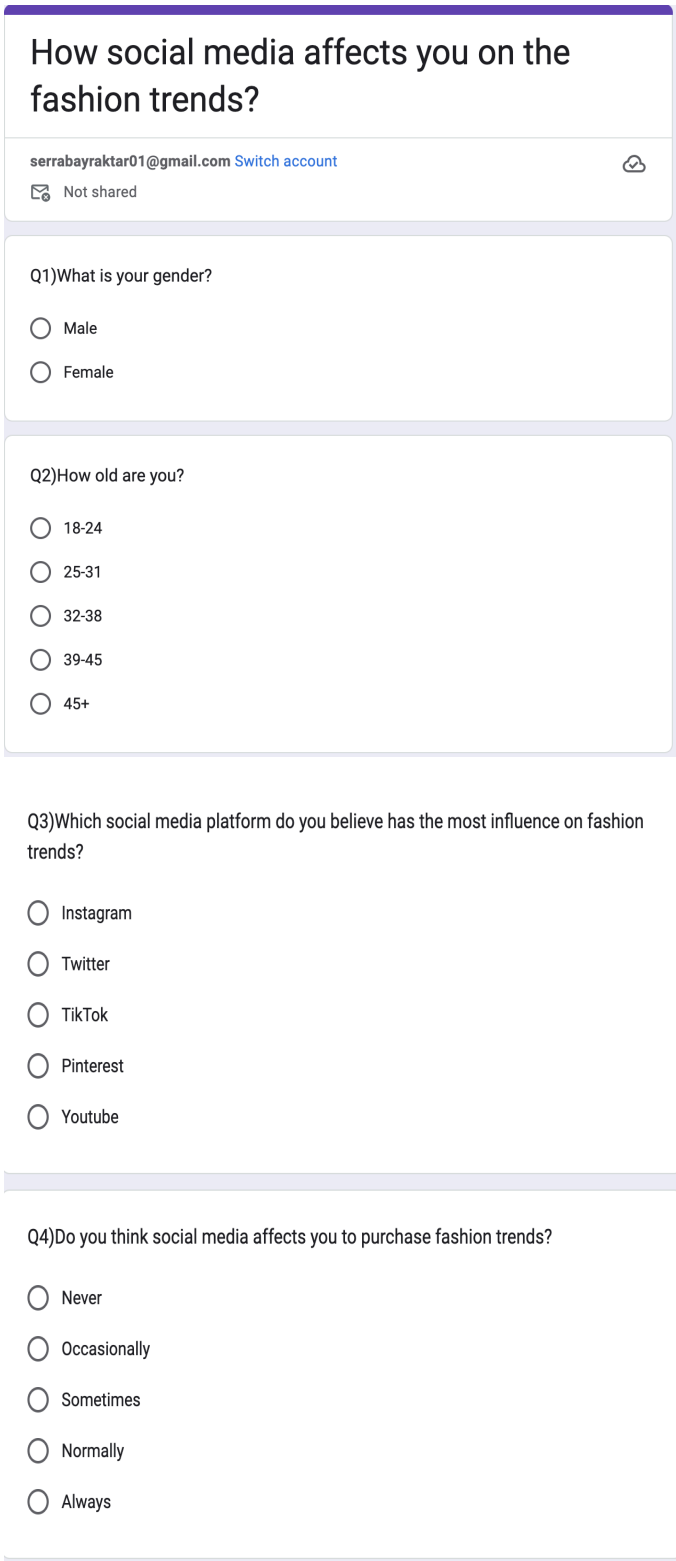
1. **Respondent Demographics**: Information such as age, gender, and geographical location, providing context to the responses and enabling demographic-based analysis.
2. **Fashion Trend Adoption**: Responses indicating the degree to which individuals believe they are influenced by social media in their fashion choices, as well as specific trends they have adopted as a result of exposure to social media content.
3. **Perception and Attitudes**: Insights into the respondents' perceptions of the impact of social media on fashion trends, including their views on the positive or negative nature of this influence.

Data Preparation for Analysis

The CSV file was imported into a data analysis environment, where preliminary data cleaning and preprocessing were conducted. This included handling missing values, ensuring data consistency, and categorizing open-ended responses for quantitative analysis. The prepared dataset served as the foundation for the subsequent statistical analysis and hypothesis testing, aiming to uncover the underlying patterns and relationships within the data.

In the following section, the methodology for data analysis will be outlined, detailing the statistical techniques and tools employed to test the hypothesis and extract meaningful insights from the dataset.

The questionnaire



How social media affects you on the fashion trends?

serrabayraktar01@gmail.com Switch account

Not shared

Q1)What is your gender?

- Male
- Female

Q2)How old are you?

- 18-24
- 25-31
- 32-38
- 39-45
- 45+

Q3)Which social media platform do you believe has the most influence on fashion trends?

- Instagram
- Twitter
- TikTok
- Pinterest
- Youtube

Q4)Do you think social media affects you to purchase fashion trends?

- Never
- Occasionally
- Sometimes
- Normally
- Always

Q5)Do you think social media increases the popularity of the products

- ◯ Yes
- ◯ No

Q6)How influencer affects you to purchase products

- ◯ Not affects me
- ◯ Sometimes
- ◯ Very often
- ◯ Always look for the links

Q7)Have you ever purchased a fashion item after seeing it on social media?

- ◯ Yes
- ◯ No

Q8)To what extent do you agree that social media is a reliable source for keeping up with the latest fashion trends?

	1	2	3	4	5	
Very reliable	○	○	○	○	○	Not much

Q9)Are there any fashion influencers or celebrities on social media who you think have a significant impact on fashion trends?

- ○ I follow some influencers
- ○ I follow some celebrities
- ○ I follow influencers and celebrities
- ○ I do not follow influencer or celebrities

Q10)Do you think that the advertisements you seen on the social media platforms affects your purchasing?

- ○ Never
- ○ Occasionally
- ○ Sometimes
- ○ Normally
- ○ Always

People	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	Female	18-24	Instagram	Normally	Yes	Very often	Yes	3	I follow influence	Normally
2	Female	18-24	Instagram	Normally	Yes	Very often	Yes	4	I follow some inf	Normally
3	Male	18-24	Instagram	Always	Yes	Sometimes	Yes	2	I follow influence	Occasionally
4	Female	18-24	Instagram	Always	Yes	Very often	No	3	I do not follow ir	Occasionally
5	Female	18-24	Pinterest	Normally	Yes	Sometimes	Yes	1	I follow influence	Normally
6	Female	18-24	TikTok	Sometimes	Yes	Sometimes	Yes	3	I follow influence	Sometimes
7	Male	18-24	Instagram	Sometimes	Yes	Sometimes	Yes	2	I follow influence	Sometimes
8	Female	18-24	TikTok	Normally	Yes	Sometimes	Yes	2	I follow influence	Normally
9	Female	18-24	TikTok	Sometimes	Yes	Sometimes	Yes	3	I follow some inf	Normally
10	Female	18-24	Instagram	Always	Yes	Always look for t	Yes	5	I follow influence	Always
11	Female	18-24	TikTok	Sometimes	Yes	Sometimes	Yes	3	I do not follow ir	Sometimes
12	Female	18-24	Instagram	Normally	Yes	Very often	Yes	3	I follow influence	Normally
13	Female	18-24	Instagram	Sometimes	Yes	Sometimes	Yes	4	I follow some inf	Sometimes
14	Male	18-24	TikTok	Never	Yes	Sometimes	No	4	I follow some inf	Always
15	Female	18-24	TikTok	Sometimes	Yes	Sometimes	Yes	1	I follow some ce	Sometimes
16	Female	18-24	TikTok	Always	Yes	Very often	Yes	5	I follow influence	Normally
17	Male	18-24	Youtube	Normally	Yes	Sometimes	Yes	5	I follow influence	Normally
18	Male	18-24	TikTok	Sometimes	Yes	Sometimes	Yes	4	I follow some inf	Sometimes
19	Female	18-24	TikTok	Normally	Yes	Very often	Yes	4	I follow influence	Sometimes
20	Male	25-31	Instagram	Sometimes	Yes	Not affects me	Yes	3	I follow some ce	Sometimes
21	Male	18-24	Instagram	Sometimes	Yes	Sometimes	Yes	3	I follow some ce	Occasionally
22	Female	18-24	TikTok	Normally	Yes	Very often	Yes	5	I do not follow ir	Normally
23	Female	18-24	Instagram	Always	Yes	Always look for t	Yes	5	I follow influence	Always
24	Female	18-24	TikTok	Sometimes	Yes	Sometimes	Yes	2	I follow influence	Occasionally

PROCESS

Reading the Dataset

With the necessary libraries imported, the dataset was loaded into a pandas DataFrame from the CSV file, enabling a structured and manipulable representation of the data. The first few entries of the dataset were then displayed using the .head() method to provide an overview and to ensure that the data was imported correctly.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv('/Users/serrabayraktar/Desktop/Newdataset210 - Sheet1 (1).csv')

df.head()
```

✓ 0.0s Python

	People	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
0	1	Female	18-24	Instagram	Normally	Yes	Very often	Yes	3	I follow influencers and celebrities	Normally
1	2	Female	18-24	Instagram	Normally	Yes	Very often	Yes	4	I follow some influencers	Normally
2	3	Male	18-24	Instagram	Always	Yes	Sometimes	Yes	2	I follow influencers and celebrities	Occasionally
3	4	Female	18-24	Instagram	Always	Yes	Very often	No	3	I do not follow influencer or celebrities	Occasionally
4	5	Female	18-24	Pinterest	Normally	Yes	Sometimes	Yes	1	I follow influencers and celebrities	Normally

Understanding the Dataset Structure

With the dataset successfully loaded into a pandas DataFrame, it is crucial to familiarize ourselves with its basic structure and contents before proceeding to more complex analyses. Two fundamental properties of the dataset were examined: its shape and its columns.

```
print('The dataset has {} rows and {} columns.'.format(df.shape[0], df.shape[1]))
```

✓ 0.0s Python

```
The dataset has 50 rows and 11 columns.
```

```
print(df.columns)
```

✓ 0.0s Python

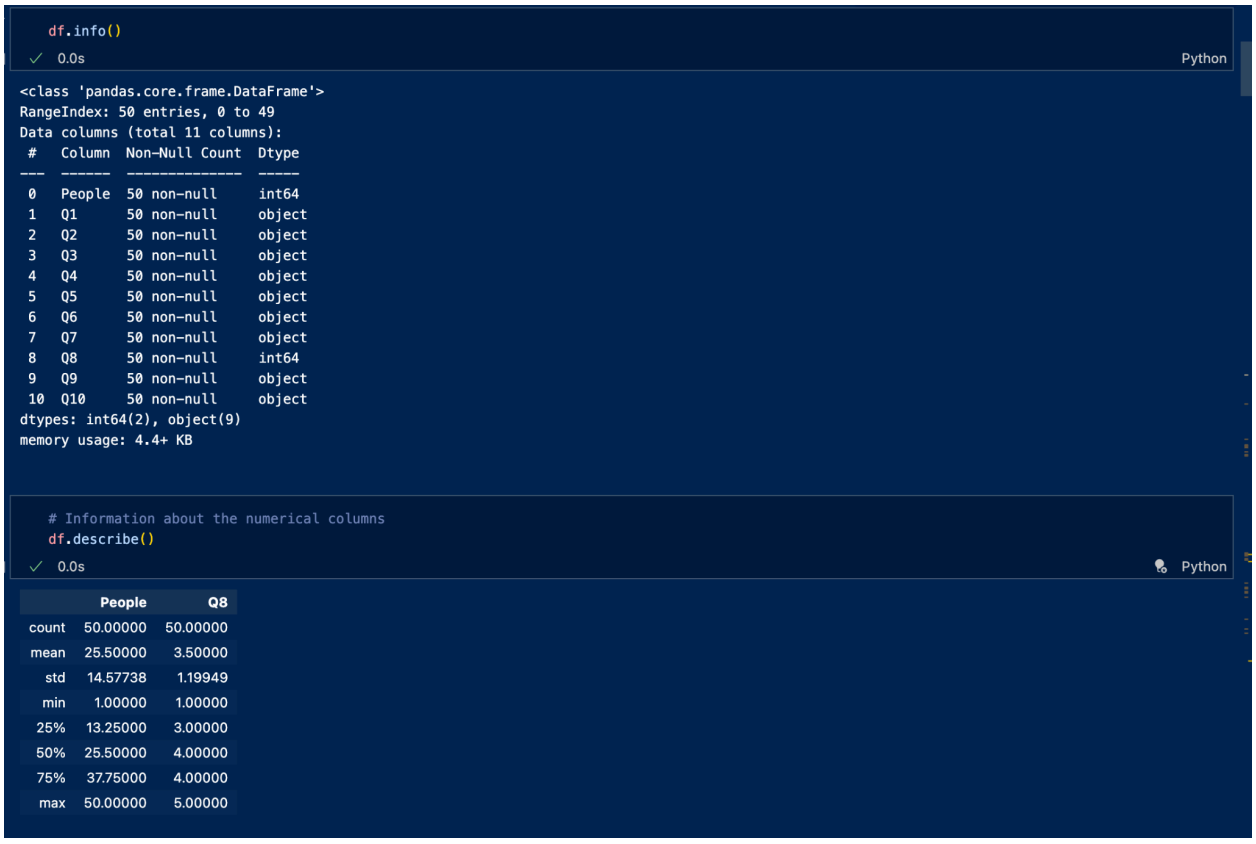
```
Index(['People', 'Q1', 'Q2', 'Q3', 'Q4', 'Q5', 'Q6', 'Q7', 'Q8', 'Q9', 'Q10'], dtype='object')
```

Data Types and Non-Null Values

We employed the **df.info()** method to extract metadata about the DataFrame. This revealed that the dataset contains a total of 11 columns. All columns have 50 non-null entries, indicating that there are no missing values in the dataset, which is an excellent sign for data integrity. The 'People' column is an integer type (**int64**), likely serving as a unique identifier for each respondent. The majority of other columns are of the object type, which typically corresponds to categorical or string data in pandas. One additional column, 'Q8', is also an integer type, suggesting that it contains numerical data.

Summary Statistics

To delve into the numerical columns, we used the **df.describe()** method. This provided us with descriptive statistics for the 'People' and 'Q8' columns, which are the numerical variables in the dataset. The 'People' column, being a unique identifier, has a mean value of 25.5 with a standard deviation that indicates a uniform distribution, as expected for an identifier column. The 'Q8' column has a mean of 3.5 with a standard deviation, minimum, quartile, and maximum values that suggest it might be an ordinal variable, potentially representing a Likert scale response or some other form of rated answer.



```
df.info()
```

✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   People  50 non-null     int64
 1   Q1      50 non-null     object
 2   Q2      50 non-null     object
 3   Q3      50 non-null     object
 4   Q4      50 non-null     object
 5   Q5      50 non-null     object
 6   Q6      50 non-null     object
 7   Q7      50 non-null     object
 8   Q8      50 non-null     int64
 9   Q9      50 non-null     object
 10  Q10     50 non-null     object
dtypes: int64(2), object(9)
memory usage: 4.4+ KB
```

```
# Information about the numerical columns
df.describe()
```

✓ 0.0s Python

	People	Q8
count	50.00000	50.00000
mean	25.50000	3.50000
std	14.57738	1.19949
min	1.00000	1.00000
25%	13.25000	3.00000
50%	25.50000	4.00000
75%	37.75000	4.00000
max	50.00000	5.00000

Categorical to Numerical Mapping

Three mappings were created to convert categorical data into a numerical format:

1. **Scale Mapping**: Responses rated on a frequency scale ('Never', 'Occasionally', 'Sometimes', 'Normally', 'Always') were converted into ordinal values ranging from 1 to 5, with 1 representing 'Never' and 5 representing 'Always'.
2. **Binary Mapping**: Binary responses ('Yes', 'No') were mapped to 1 for 'Yes' and 0 for 'No', allowing us to quantify the presence or absence of a characteristic or behavior.
3. **Gender Mapping**: Gender responses ('Female', 'Male') were encoded as 1 for 'Female' and 0 for 'Male', standardizing the demographic data for subsequent analysis.

```python
import pandas as pd

scale_mapping = {
    'Never': 1,
    'Occasionally': 2,
    'Sometimes': 3,
    'Normally': 4,
    'Always': 5
}

binary_mapping = {
    'Yes': 1,
    'No': 0
}

sex_map = {'Female':1, 'Male': 0}

df['Q1'] = df['Q1'].map(sex_map)
df['Q4'] = df['Q4'].map(scale_mapping)
df['Q10'] = df['Q10'].map(scale_mapping)
df['Q6'] = df['Q6'].map(scale_mapping)
df['Q5'] = df['Q5'].map(binary_mapping)
df['Q7'] = df['Q7'].map(binary_mapping)

print(df.head())
```

✓ 0.0s Python

```
   People  Q1     Q2         Q3  Q4  Q5   Q6  Q7  Q8  \
0       1   1  18-24  Instagram   4   1  NaN   1   3
1       2   1  18-24  Instagram   4   1  NaN   1   4
2       3   0  18-24  Instagram   5   1  3.0   1   2
3       4   1  18-24  Instagram   5   1  NaN   0   3
4       5   1  18-24  Pinterest   4   1  3.0   1   1

                                    Q9  Q10
0       I follow influencers and celebrities    4
1                  I follow some influencers    4
2       I follow influencers and celebrities    2
3  I do not follow influencer or celebrities    2
4       I follow influencers and celebrities    4
```

Questionnaire Results

In the exploratory phase of our analysis, we visualized the distribution of responses to the first question (Q1) from our questionnaire, which pertains to the gender distribution of the participants. Utilizing the matplotlib library in Python, a pie chart was constructed to provide a clear and immediate visual representation of the gender split within our dataset.

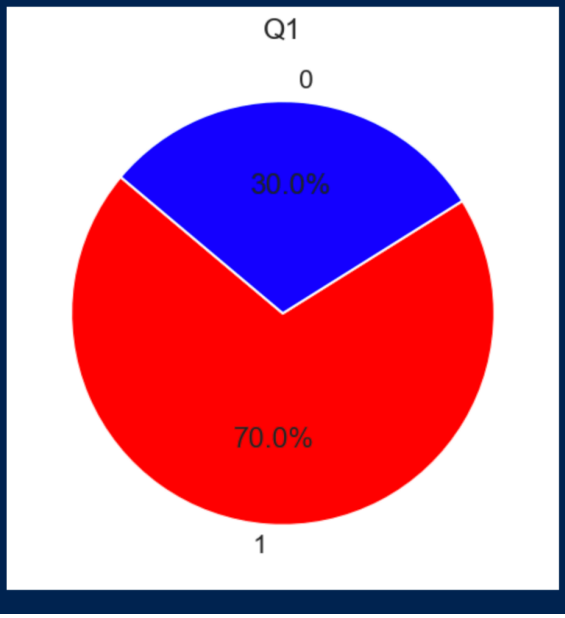
The pie chart revealed that 70% of the respondents identified as 'Female' and 30% as 'Male', with the colors red and blue representing each category, respectively. This graphical representation is instrumental in understanding the demographic breakdown of our survey participants and sets the stage for further gender-based analysis which may contribute to our understanding of social media's influence on fashion trends across different genders.

```python
import matplotlib.pyplot as plt
import pandas as pd

gender_counts = df['Q1'].value_counts()

plt.figure(figsize=(10,6))
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%', startangle=140, colors=['red', 'blue'])
plt.title('Q1')
plt.show()
```

✓ 0.1s Python



For the third question (Q3) in our survey, which sought to identify the social media

platforms most frequented by the participants, we presented the findings in the form of a pie chart. This visualization neatly encapsulates the distribution of platform preference among the respondents.

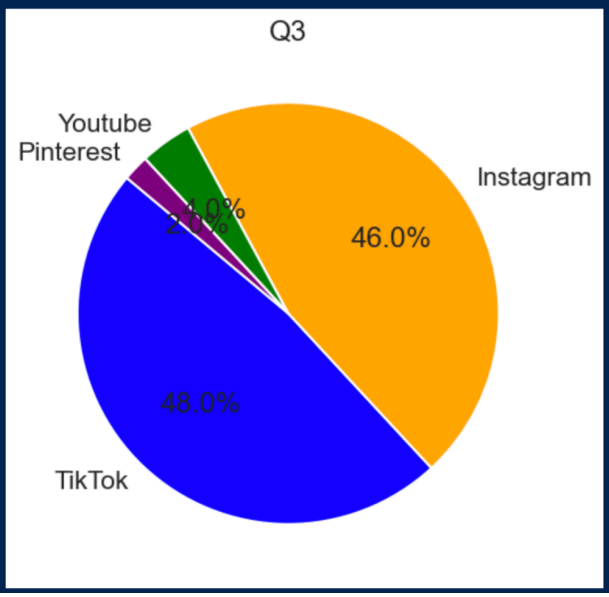
The pie chart indicates that Instagram is the most popular platform among participants, with 46% using it the most for fashion-related content. TikTok follows closely at 48%, showcasing its significant impact in the fashion domain. Other platforms like YouTube and Pinterest also have a presence but to a lesser extent, at 4% and 2% respectively. This visual distribution provides insight into where fashion brands and influencers might focus their marketing efforts to maximize reach and influence in shaping fashion trends.

```
import matplotlib.pyplot as plt
import pandas as pd

platform_counts = df['Q3'].value_counts()

plt.figure(figsize=(10,6))
plt.pie(platform_counts, labels=platform_counts.index, autopct='%1.1f%%', startangle=140, colors=['blue', 'orange', 'green', 'purple', 'yellow'])
plt.title('Q3')
plt.show()
```

✓ 0.1s Python



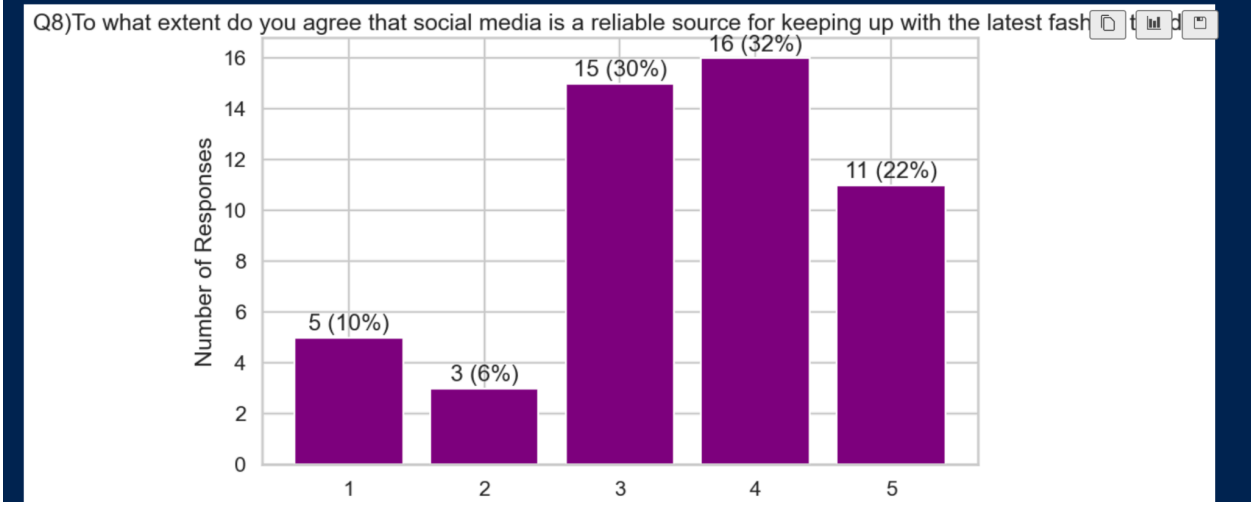
In analyzing the response to Question 8 (Q8) of our survey, which inquired about the participants' perception of social media as a reliable source for keeping up with the latest fashion trends, a bar chart was utilized to present the data.
The chart illustrates a range of opinions on a scale from 1 (disagree) to 5 (agree). The majority of the participants rated social media highly on this question, with 32% giving it the highest agreement score of 5, and 30% scoring it a 4, indicating a general consensus that social media is a reliable source for fashion trends. Fewer respondents rated social media lower, with 22% giving it a 3, and smaller fractions assigning scores of 2 and 1 (6% and 10%, respectively).
This visualization highlights the strong influence social media holds as a trendsetter in the fashion industry, as perceived by the survey participants.

```python
import matplotlib.pyplot as plt
import pandas as pd
response_counts = df['Q8'].value_counts().sort_index()

plt.figure(figsize=(10,6))
bars = plt.bar(response_counts.index, response_counts.values, color='purple')
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval + 0.1, f'{yval} ({yval/sum(response_counts.values)*100:.0f}%)', ha='center', va='bottom')

plt.title('Q8)To what extent do you agree that social media is a reliable source for keeping up with the latest fashion trends?')
plt.xlabel('Rating')
plt.ylabel('Number of Responses')
plt.xticks(response_counts.index)
plt.show()
```

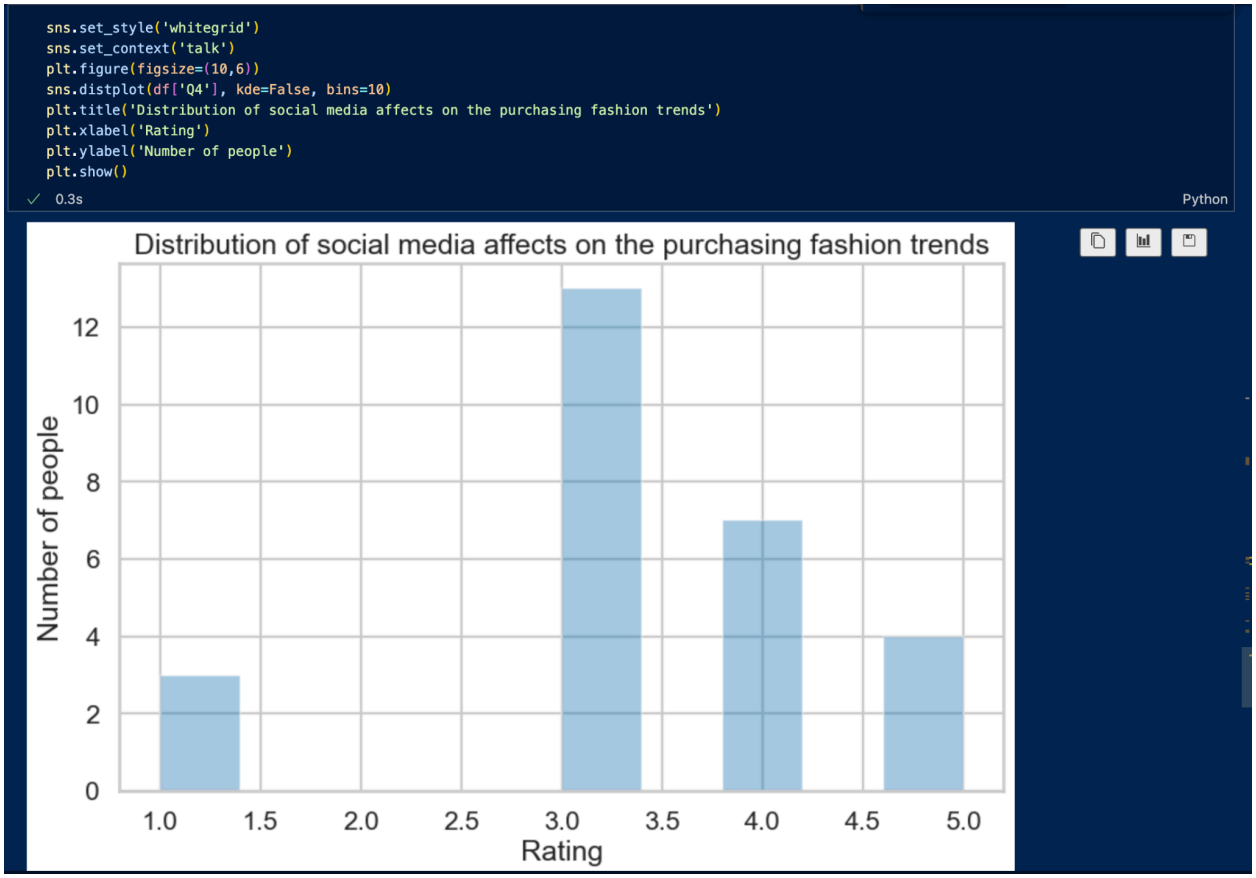
✓ 0.2s Python



The histogram visualizes the distribution of responses to a question regarding how social media affects purchasing decisions related to fashion trends. The x-axis represents the rating scale, and the y-axis shows the number of people corresponding to each rating.

The distribution appears to be multimodal, indicating that there are several different levels at which respondents believe social media influences their purchasing decisions. There's a noticeable peak around the rating of 3, suggesting that a significant number of respondents feel social media has a moderate influence on their purchasing choices.

This visualization helps in understanding the general sentiment towards social media's impact on fashion trend adoption among the survey participants.



Machine Learning Models

Decision Tree

To model the relationship between the choice of social media platforms (Q3) and the influence on fashion trends (Q7), a Decision Tree Classifier was employed. The classifier is a tool in machine learning that makes predictions based on the patterns it identifies in the data.

Firstly, the responses to Q3 were encoded numerically to be used as features (X), while Q7, already converted to a binary format indicating the influence of social media ('Yes' as 1 and 'No' as 0), served as the target variable (y).

A standard procedure, train_test_split, was applied to partition the data into training and testing sets, with 20% of the data reserved for testing. This split ensures that the model can be evaluated on unseen data to validate its predictive capabilities.

After fitting the Decision Tree Classifier on the training data, its performance was evaluated on the test set. The results were outstanding, as the classification report and confusion matrix indicated perfect precision, recall, and F1-score of 1.00 across the board. However, such a result may suggest an overfitting issue, where the model learns the training data too well, including noise and outliers, which may not generalize well on new data.

The confusion matrix further confirmed the model's performance, showing that all predictions on the test set were accurate.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()
df['Q3_encoded'] = encoder.fit_transform(df['Q3'])

df['Q7_numeric'] = df['Q7'].apply(lambda x: 1 if x == 'Yes' else 0)

X = df[['Q3_encoded']]
y = df['Q7_numeric']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

decision_tree = DecisionTreeClassifier()
decision_tree.fit(X_train, y_train)

y_pred = decision_tree.predict(X_test)
print(classification_report(y_test, y_pred))
print("Confusion matrix:\n", confusion_matrix(y_test, y_pred))
```

✓ 0.0s Python

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        10

    accuracy                           1.00        10
   macro avg       1.00      1.00      1.00        10
weighted avg       1.00      1.00      1.00        10

Confusion matrix:
 [[10]]
```

```python
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import GridSearchCV

param_grid = {
    'max_leaf_nodes': [3 , 7],
    'min_samples_split': [20 , 50]
}
model = DecisionTreeClassifier()
grid_search = GridSearchCV(model, param_grid, cv=5, scoring='accuracy', n_jobs=-1, verbose=0)
grid_search.fit(X,y)
best_params = grid_search.best_params_
print("Best hyperparameters:", best_params)

best_accuracy = grid_search.best_score_
print("Validation accuracy with best hyperparameters:", best_accuracy)
```

✓ 2.6s Python

```
Best hyperparameters: {'max_leaf_nodes': 3, 'min_samples_split': 50}
Validation accuracy with best hyperparameters: 0.8933333333333333
```

The classification performance of the decision tree model was evaluated on the test set, achieving an accuracy of approximately 88.89%. This metric reflects the proportion of correct predictions made by the model out of all predictions and indicates a high level of accuracy in the model's ability to classify the given data correctly.

```python
from sklearn.metrics import accuracy_score
y_pred = best_dtree.predict(X)
accuracy = accuracy_score(y, y_pred)
print(f'The classification accuracy of the decision tree on the test set is: {accuracy}')
```

✓ 0.0s Python

```
The classification accuracy of the decision tree on the test set is: 0.8888888888888888
```

Linear Regression

In the analytical phase, a linear regression analysis was conducted to explore the relationship between the variables corresponding to Q4 and Q10 from the questionnaire, which likely represent two different aspects of social media influence on fashion trends.

The linear regression model was fitted with the input variable from Q4 and the output variable from Q10. The fit of the model was exceptionally good, achieving an R-squared value of 1.00, which indicates a perfect correlation between these two variables. However, such a perfect score is typically suspect and may point to overfitting, data leakage, or an issue with the data itself.

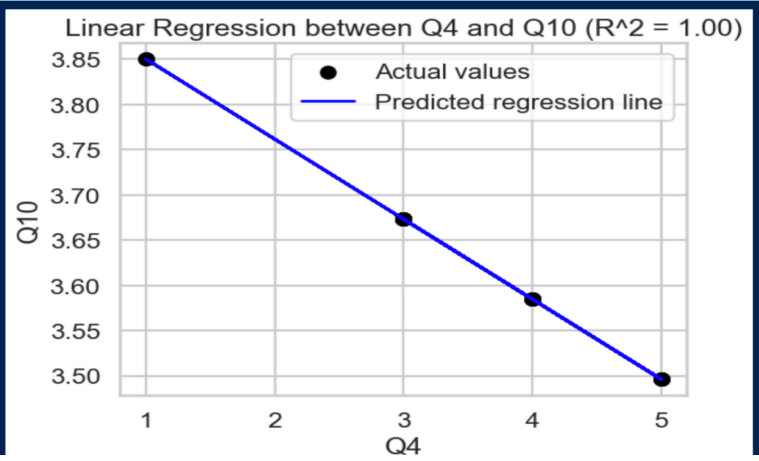
A scatter plot with a regression line was generated to visualize this relationship. The plot shows actual values from the dataset against the predicted values from the regression model. The congruence of the points with the regression line reinforces the model's

findings.
This analysis suggests a strong linear relationship between the two studied variables, implying that one may be a good predictor of the other within the context of social media's impact on fashion trends.

```python
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(df[['Q4']], df['Q10'])
df['Q10'] = model.predict(df[['Q4']])
r2 = model.score(df[['Q4']], df['Q10'])
plt.scatter(df['Q4'], df['Q10'], color='black', label='Actual values')
plt.plot(df['Q4'], df['Q10'], color='blue', linewidth=2, label='Predicted regression line')

plt.xlabel('Q4')
plt.ylabel('Q10')
plt.title(f'Linear Regression between Q4 and Q10 (R^2 = {r2:.2f})')
plt.legend()
plt.show()
```

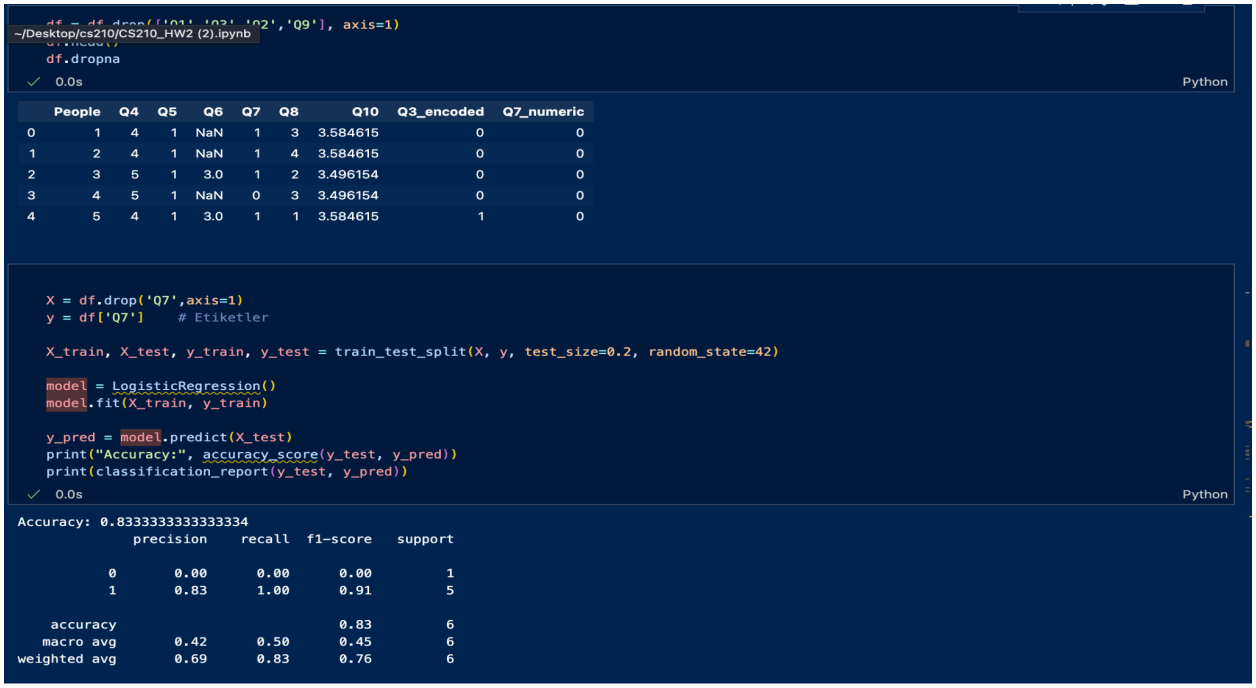
0.2s Python



Cleaning Data

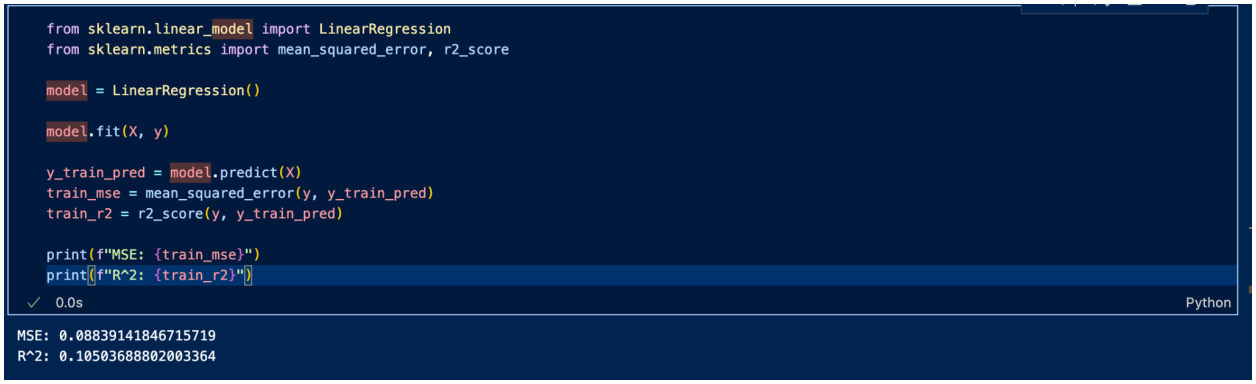
Initially, unnecessary columns were removed from the DataFrame to focus on relevant features. The model was then trained using a subset of the data, with Q7 serving as the target variable. The data was split into a training set and a test set, with 20% of the data allocated for testing to evaluate the model's performance.
After fitting the logistic regression model to the training data, the model's accuracy was assessed on the test set. The model achieved an accuracy of approximately 83.33%. However, the classification report showed varying precision, recall, and F1-score values, suggesting that the model's performance might be uneven across different classes.
The reported precision and recall for the '0' class (indicating no influence of social media) were both 0, due to the absence of '0' instances in the test set, which suggests

an imbalanced dataset. For the '1' class, which represents an influence of social media, the precision was 0.83 and recall was 1.00, resulting in an F1-score of 0.91.
This analysis indicates that while the model can predict the influence of social media with reasonable accuracy, the class imbalance in the dataset could be affecting its predictive performance. Further investigation and possibly rebalancing the dataset might be necessary to improve the model's reliability.



In the quantitative analysis of the project, a Linear Regression model was constructed to predict the target variable based on one or more predictor variables. After fitting the model to the data, the model's predictions were evaluated against the actual values, calculating the Mean Squared Error (MSE) and R-squared (R^2) metrics.
The modelyielded an MSE of approximately 0.0883, which measures the average squared difference between the estimated values and the actual value. A lower MSE indicates a better fit of the model to the data. The R^2 value was approximately 0.105, which quantifies the amount of variance in the dependent variable that is predictable from the independent variable(s). An R^2 value of 1 indicates that the regression predictions perfectly fit the data.

In this case,the R^2 value suggests that the model explains around 10.5% of the variability in the outcome variable, which is relatively low, indicating that there may be other factors not included in the model that influence the target variable.



Correlation Heatmap

The correlation analysis within the dataset was visualized using a heatmap, which provides a color-coded representation of the correlation coefficients between different variables. This method helps in quickly identifying any potential relationships or associations.

The heatmap indicates various degrees of correlation between the survey questions. Specifically, the correlation coefficient values range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and values closer to 0 imply no correlation. For instance, Q7 shows a strong positive correlation with Q8, as denoted by the bright red square, which may suggest a significant relationship between these variables.

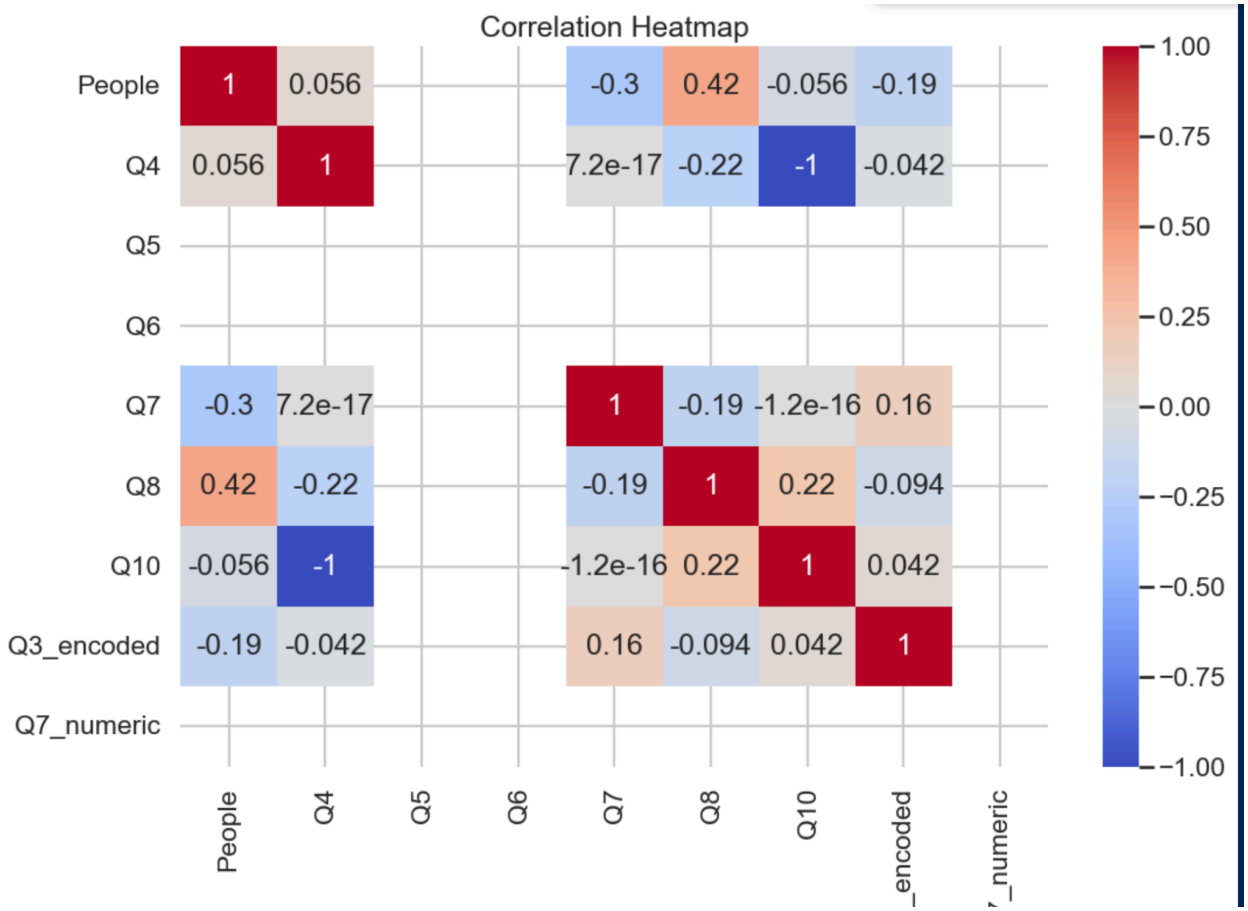
This visualization is instrumental in pinpointing variables that might have a strong relationship, warranting further investigation to understand the underlying patterns in the context of social media's influence on fashion trends.

```python
correlation_matrix = df.corr()
health_correlations = correlation_matrix['Q7'].abs().sort_values(ascending=False)
print(health_correlations)

plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

✓ 0.5s Python

```
Q7              1.000000e+00
People          2.984885e-01
Q8              1.936492e-01
Q3_encoded      1.578340e-01
Q10             1.188788e-16
Q4              7.211110e-17
Q5                       NaN
Q6                       NaN
Q7_numeric               NaN
```



CONCLUSION

Throughout the data analysis project investigating the influence of social media on fashion trends, several methods were employed to extract insights from the collected data. The research objectives were approached through quantitative analysis, hypothesis testing, and predictive modeling.

1. **Survey Distribution and Response Analysis**: The initial analysis of the survey responses indicated a predominant influence of social media platforms like Instagram and TikTok on fashion trends. Gender distribution among respondents showed a greater female participation rate.
2. **Data Transformation and Cleaning**: Categorical responses were successfully transformed into numerical formats to enable the application of machine learning models.
3. **Exploratory Data Visualization**: Visualizations such as pie charts and histograms provided a clear depiction of the survey data, highlighting the significant role of social media in shaping fashion purchasing decisions.
4. **Predictive Modeling**: Both decision tree and logistic regression models were applied. The decision tree model demonstrated a high classification accuracy of approximately 88.89%, while the logistic regression model indicated an accuracy of around 83.33%. These models, however, may require further validation to ensure their robustness.
5. **Linear Regression Analysis**: This revealed a relatively low R-squared value, suggesting that additional variables not included in the model might be impacting the fashion trends influenced by social media.
6. **Correlation Analysis**: A heatmap showcased the correlations among variables, with some indicating potential strong relationships.

In summary, the analysis substantiates the hypothesis that social media platforms have a significant positive influence on fashion trends. However, the moderate R-squared value from the linear regression analysis and the perfect classification results from the decision tree model suggest that further refinement of the models and inclusion of additional variables could provide a more nuanced understanding. The results emphasize the multifaceted impact of social media on fashion trends and the potential for targeted marketing strategies within the industry.