

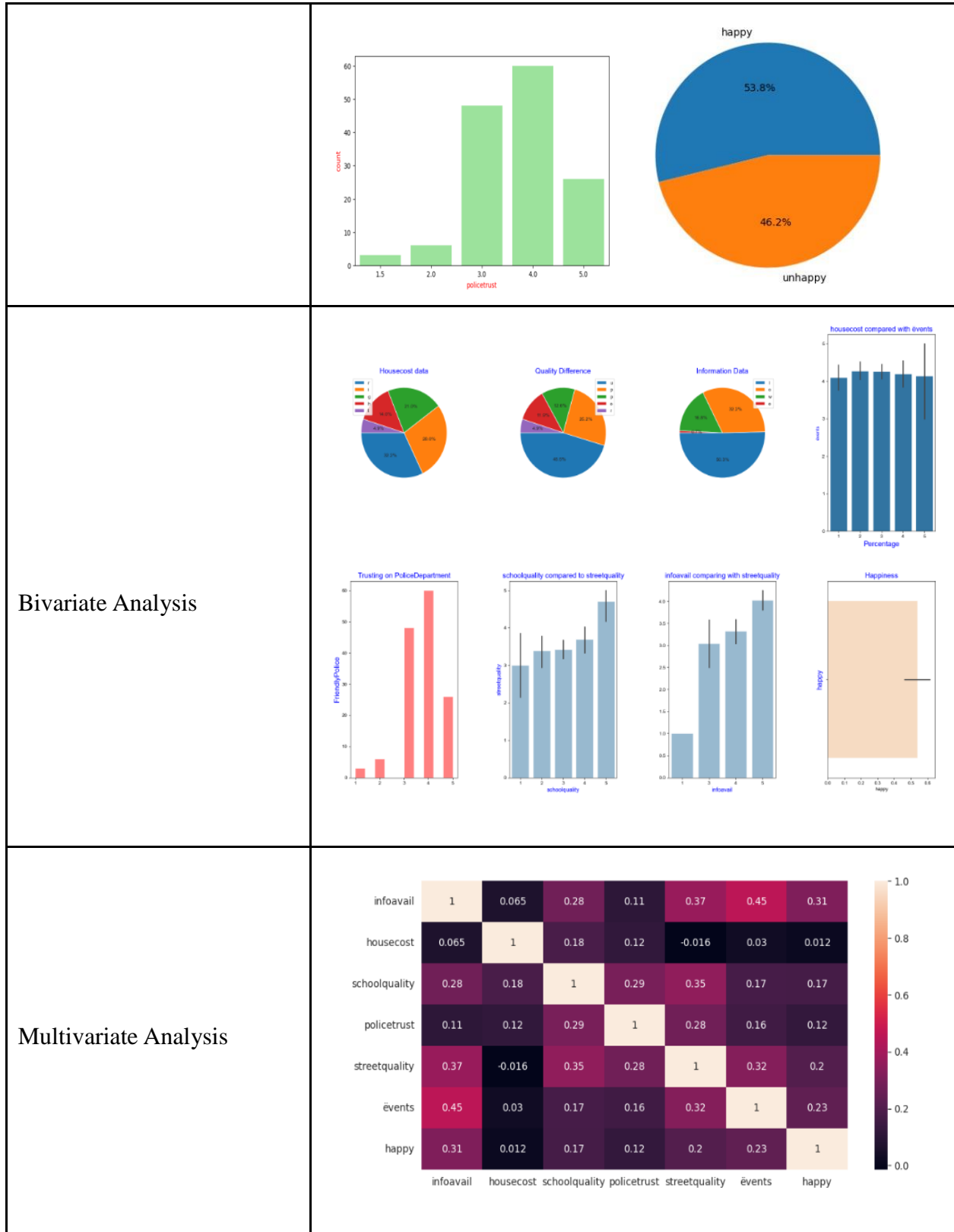
Data Collection and Preprocessing Phase

Date	21 June 2024
Team ID	739931
Project Title	Eudaimonia Engine: Machine Learning Delving into Happiness Classification
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description																																																																								
Data Overview	<u>Dimension:</u> <u>143 rows \times 7 columns</u> <u>Descriptive statistics:</u>																																																																								
	<table><thead><tr><th></th><th>infoavail</th><th>housecost</th><th>schoolquality</th><th>policetrust</th><th>streetquality</th><th>events</th><th>happy</th></tr></thead><tbody><tr><td>count</td><td>143.000000</td><td>143.000000</td><td>143.000000</td><td>143.000000</td><td>143.000000</td><td>143.000000</td><td>143.000000</td></tr><tr><td>mean</td><td>4.325175</td><td>2.513986</td><td>3.265734</td><td>3.699301</td><td>3.615385</td><td>4.216783</td><td>0.538462</td></tr><tr><td>std</td><td>0.765126</td><td>1.068011</td><td>0.992586</td><td>0.888383</td><td>1.131639</td><td>0.848693</td><td>0.500271</td></tr><tr><td>min</td><td>2.500000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>0.000000</td></tr><tr><td>25%</td><td>4.000000</td><td>2.000000</td><td>3.000000</td><td>3.000000</td><td>3.000000</td><td>4.000000</td><td>0.000000</td></tr><tr><td>50%</td><td>5.000000</td><td>3.000000</td><td>3.000000</td><td>4.000000</td><td>4.000000</td><td>4.000000</td><td>1.000000</td></tr><tr><td>75%</td><td>5.000000</td><td>3.000000</td><td>4.000000</td><td>4.000000</td><td>4.000000</td><td>5.000000</td><td>1.000000</td></tr><tr><td>max</td><td>5.000000</td><td>4.500000</td><td>5.000000</td><td>5.000000</td><td>5.000000</td><td>5.000000</td><td>1.000000</td></tr></tbody></table>		infoavail	housecost	schoolquality	policetrust	streetquality	events	happy	count	143.000000	143.000000	143.000000	143.000000	143.000000	143.000000	143.000000	mean	4.325175	2.513986	3.265734	3.699301	3.615385	4.216783	0.538462	std	0.765126	1.068011	0.992586	0.888383	1.131639	0.848693	0.500271	min	2.500000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	25%	4.000000	2.000000	3.000000	3.000000	3.000000	4.000000	0.000000	50%	5.000000	3.000000	3.000000	4.000000	4.000000	4.000000	1.000000	75%	5.000000	3.000000	4.000000	4.000000	4.000000	5.000000	1.000000	max	5.000000	4.500000	5.000000	5.000000	5.000000	5.000000	1.000000
		infoavail	housecost	schoolquality	policetrust	streetquality	events	happy																																																																	
	count	143.000000	143.000000	143.000000	143.000000	143.000000	143.000000	143.000000																																																																	
	mean	4.325175	2.513986	3.265734	3.699301	3.615385	4.216783	0.538462																																																																	
	std	0.765126	1.068011	0.992586	0.888383	1.131639	0.848693	0.500271																																																																	
	min	2.500000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000																																																																	
	25%	4.000000	2.000000	3.000000	3.000000	3.000000	4.000000	0.000000																																																																	
	50%	5.000000	3.000000	3.000000	4.000000	4.000000	4.000000	1.000000																																																																	
	75%	5.000000	3.000000	4.000000	4.000000	4.000000	5.000000	1.000000																																																																	
max	5.000000	4.500000	5.000000	5.000000	5.000000	5.000000	1.000000																																																																		
Univariate Analysis																																																																									



Outliers and Anomalies



Data Preprocessing Code Screenshots

Loading Data

```
[ ] #READ THE DATASET

df=pd.read_csv("/content/happydata.csv")
df.head( )
```

	infoavail	housecost	schoolquality	policetrust	streetquality	events	happy
0	3	3	3	4	2	4	0
1	3	2	3	5	4	3	0
2	5	3	3	3	3	5	1
3	5	4	3	3	3	5	0
4	5	4	3	3	3	5	0

Handling Missing Data	<pre>#DATA PREPARATION #HANDLING MISSING VALUES df.isnull().any()</pre> <pre>infoavail False housecost False schoolquality False policetrust False streetquality False events False happy False dtype: bool</pre> <pre>[] df.isnull().sum()</pre> <pre>infoavail 0 housecost 0 schoolquality 0 policetrust 0 streetquality 0 events 0 happy 0 dtype: int64</pre> <pre>#HANDLING DUPLICATES VALUES df.duplicated().sum()</pre> <pre>18</pre> <pre>[] df.drop_duplicates()</pre> <pre>infoavail housecost schoolquality policetrust streetquality events happy 0 3 3 3 4 2 4 0 1 3 2 3 5 4 3 0 2 5 3 3 3 3 5 1 3 5 4 3 3 3 5 0 5 5 5 3 5 5 5 1 137 5 2 3 4 4 3 1 138 5 3 3 1 3 5 0 139 5 2 3 4 2 5 1 141 4 3 3 4 4 5 0 142 5 3 2 5 5 5 0</pre> <p>125 rows x 7 columns</p>
Data Transformation	<pre>[] # Separate the independent variables x = df.drop(columns='happy',axis=1) # Separate the target variable y = df['happy'] from sklearn.model_selection import train_test_split x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.2, random_state=0)</pre>
Feature Engineering	Attached are the codes in the final submission.
Save Processed Data	-