Flourishing dataset analysis

Fulin ZHANG

I. Introduction

This was a personal project exploring the Flourishing dataset, where I progressively performed data mining, supervised learning, and unsupervised learning. Here is a short report on my exploration of the dataset.

II. Data exploring

Univariate analysis:
I describe the distribution firstly. I got no missing value. And after analyzing the histograms and box plots of the quantitative features, I found some interesting information. And here is some description for **quantitive features**.

> **age**: The mean age was 41.6 and the distribution began at age 20 and continued until age 72. The overall leftward skew indicates that the majority of the sample subjects are young adults.
> **education**: Education is a discrete variable, and the data are significantly right-skewed, indicating that there is a large sample of highly educated individuals.
> **priv_quant & pro_quant**: Because of the small sample size, I can only show one normal distribution for these two samples roughly.
> **positivity**: The mean is 1.8, with most of the distribution in the 1 to 2 range.

I then manipulated the **categorical variables** with pie charts and bar charts to see the distribution of the share and number of each category.

> **sex**: The proportion of women is 60.9%, showing an unbalanced distribution of more women than men.
> **famstatus** : Couples accounted for 69.8% of family status.
> **pro_cat** : The professional context Moderately Mentally Healthy has the highest percentage, occupying 64.1%, while the other two are roughly evenly distributed.
> **priv_cat** : Private Context Moderately Mentally Healthy still has the largest share, 66.1%. languishing is the least, occupying 5.6%.

Bivariate analysis

After exploring the relationship between the **quantitative variables and the target values**, it was found that pro_quant had a strong positive correlation with it, followed by priv_quant and positivity, and a sizable positive correlation with age. education had a somewhat negative correlation. (Figure 1)

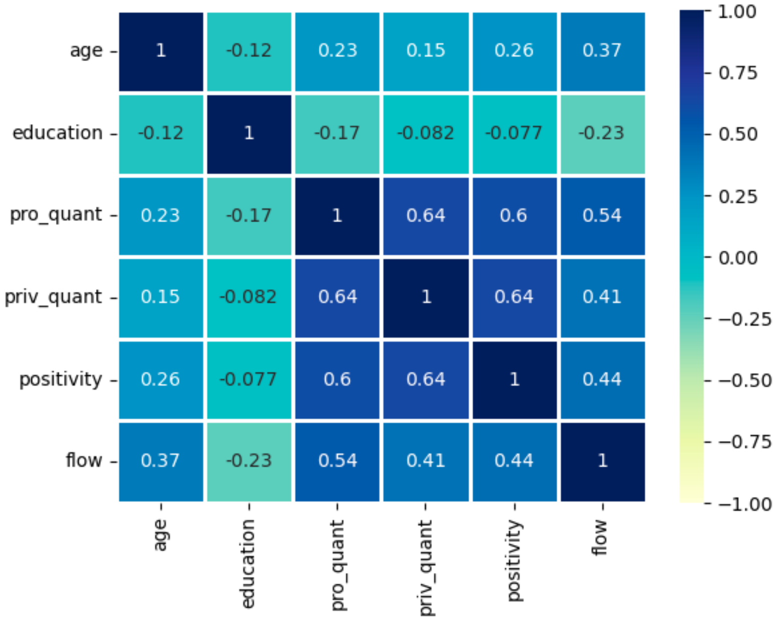
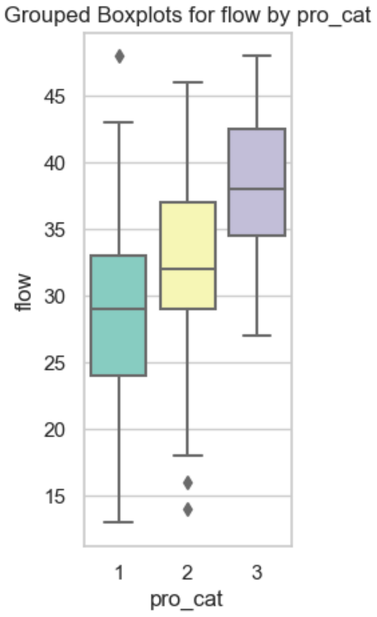


Figure 1

I then explored the relationship between **categorical and target** variables by ANOVA and came to the following conclusions.

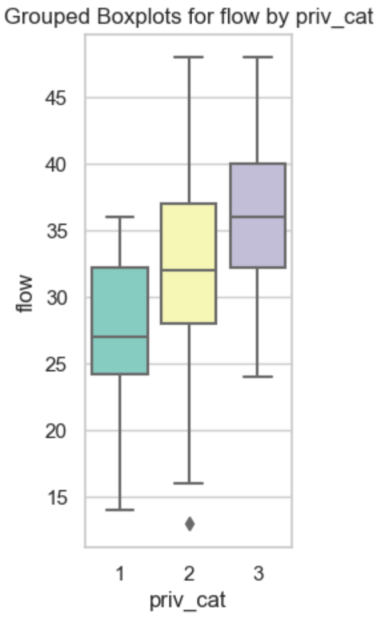
pro_cat – flow : p-value = 1.073305e-11 << 0.05

Significant correlation with target value

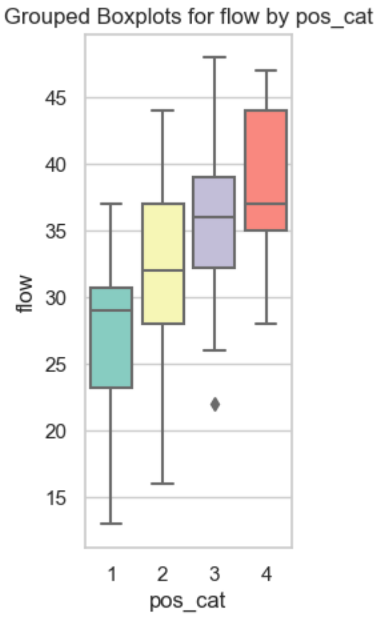


pro_cat – flow : p-value = 1.432929e-07 << 0.05

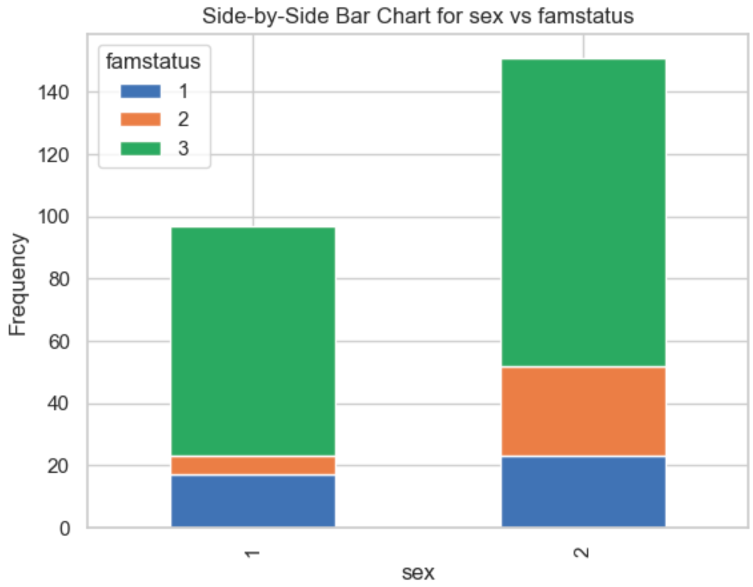
Significant correlation with target value



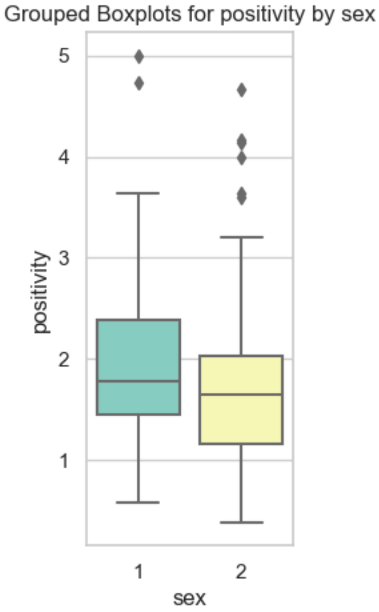
pos_cat – flow : p-value = 7.450157e-12 << 0.05
Significant correlation with target value



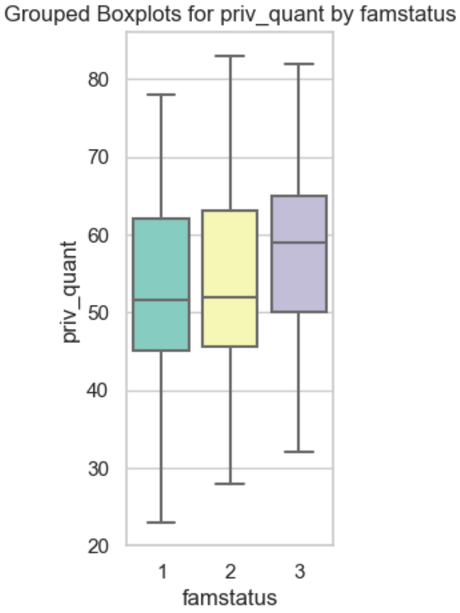
sex_cat – flow : p-value = 0.48745 >> 0.05
No significant difference from target
So I ran a chi-square test on SEX and FAMSTATUS and found that there is some association between them, p-value = 0.01607817439257878



I then examined the relationship between SEX and POSITIVITY and obtained a p-value of 0.036316, indicating that the original hypothesis could be discarded. The positivity of 1 (male) in the sample was generally slightly higher than that of 2 (female).



sex_cat – flow : p-value = 0.186796 >> 0.05
No significant difference from target
Then since it's a family, it could be related to ratings in private status, so I did correlation mining between priv_quant and famstatus. I got a p-value of 0.022079, so again the original hypothesis can be discarded. Priv_quant is relatively high for overall COUPLES in the sample and lowest for SINGLES.



III. Supervised – quantitive target

After the previous analysis, the data were all inextricably linked to each other, so I kept them and then performed supervised learning, using a target value of the quantitative variable flow. LinearRegression and KNeighborsRegressor models were used, respectively. The mse and r^2 were applied to evaluate the models and the following are the evaluation results obtained.

```
lr train_mse : 23.282429612542302
lr test_mse : 34.2519688214407
lr train_r2 : 0.4663195363849406
lr test_r2 : 0.19140772376202297

KNN train_mse : 24.798383838383838
KNN test_mse : 31.383999999999997
KNN train_r2 : 0.43157079376958707
KNN test_r2 : 0.2591123701605287
```

For **Linear Regression,** the MSE (Mean Squared Error) on the training set is low at about 23.28 while the MSE on the test set is high at 34.25.This may indicate that the model fits better on the training set but has poor generalization performance on the unseen data.

The R^2 value of 0.4663 on the training set indicates that the model is able to explain about 46.63% of the variance in the data, whereas the R^2 value on the test set is lower at 0.1914, which implies that the model's ability to explain on the unseen data is reduced.

For **K Nearest Neighbor** Analysis.

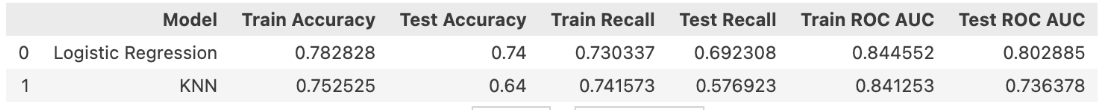
The MSE on the training set is slightly higher at around 24.80, but the MSE on the test

set is 31.38, which shows that KNN performs slightly better on the test set as compared to the linear regression model.
The R^2 value of KNN model on the training set is 0.4316 which is similar to linear regression but on the test set it is 0.2591 which is significantly better than linear regression.

So comparing the two models together, linear regression slightly outperforms KNN on the training data, but on the test data, KNN performs better and shows better generalization. This may indicate that the linear regression model overfitted on this dataset, while the KNN model better captures the intrinsic structure of the data. In practical applications, we may tend to choose KNN as the preferred model considering the importance of generalization ability.

IV. Supervised – categorical target

Here I classify the quantitative variable flow into two groups according to its median and then apply logistic regression and knn classifier to classify them respectively. After running, we obtain the following evaluation metrics.



	Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Logistic Regression	0.782828	0.74	0.730337	0.692308	0.844552	0.802885
1	KNN	0.752525	0.64	0.741573	0.576923	0.841253	0.736378

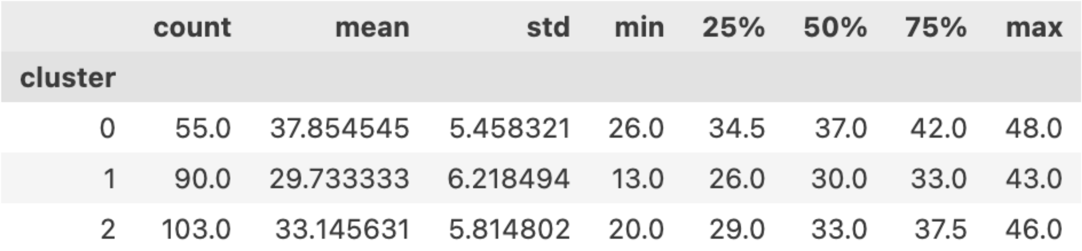
For **logistic regression**, on the training set: the accuracy reached 78.28%, the recall was 73.03%, and the ROC AUC score was 84.45%. These metrics indicate that the model has good performance on the training data. On the test set: the accuracy is 74%, the recall is 69.23%, and the ROC AUC score is 80.29%. Compared to the training set, these metrics are slightly decreased but still high, indicating that the model has good generalization ability.

For **KNN**, on the training set: accuracy is 75.25%, recall is 74.16%, and ROC AUC score is 84.13%. These numbers are slightly lower compared to logistic regression.
On the test set: accuracy is 64%, recall is 57.69%, and ROC AUC score is 73.64%. These metrics are a large drop from those on the training set, suggesting that the KNN may be somewhat overfitted on this dataset.

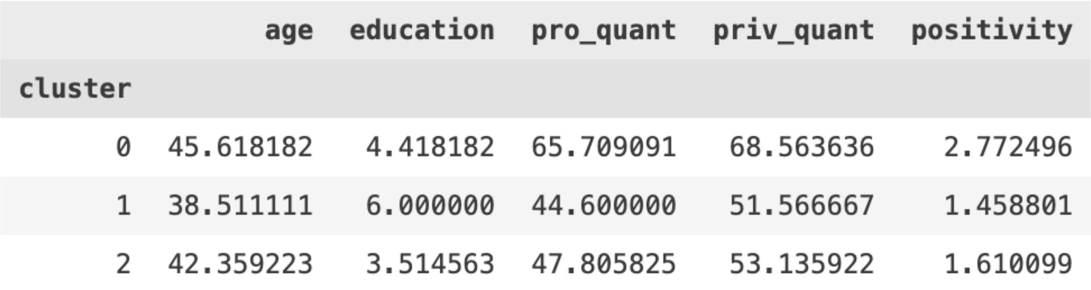
Therefore, both logistic regression and KNN perform well on the training data, but logistic regression significantly outperforms KNN on the test data. This may indicate that the logistic regression model is more stable on this dataset and is better able to capture the intrinsic structure of the data. Therefore, we may be inclined to choose logistic regression as the preferred model.

IV. Unsupervised

For unsupervised learning, the essence is to let it autonomously find the same points and cluster itself based on the distance, here I set it to 3 clusters, and then the following is the data derived from the categorization, and I explored the other variables for each cluster separately and then came up with a rough statistic.



	count	mean	std	min	25%	50%	75%	max
cluster								
0	55.0	37.854545	5.458321	26.0	34.5	37.0	42.0	48.0
1	90.0	29.733333	6.218494	13.0	26.0	30.0	33.0	43.0
2	103.0	33.145631	5.814802	20.0	29.0	33.0	37.5	46.0



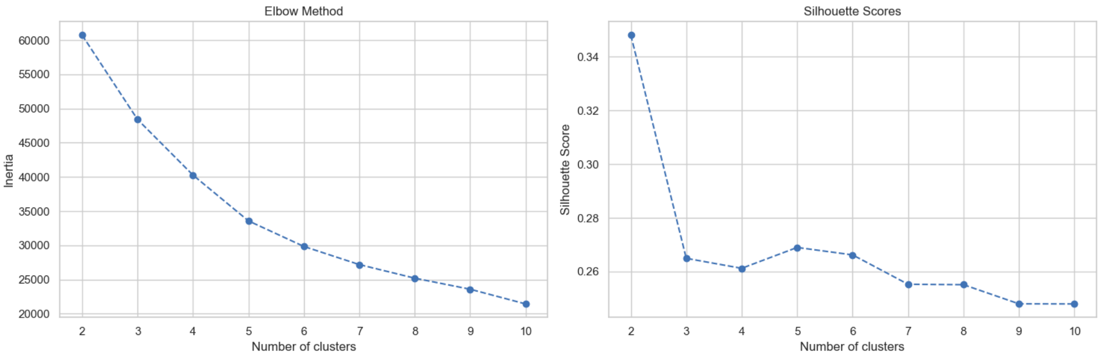
	age	education	pro_quant	priv_quant	positivity
cluster					
0	45.618182	4.418182	65.709091	68.563636	2.772496
1	38.511111	6.000000	44.600000	51.566667	1.458801
2	42.359223	3.514563	47.805825	53.135922	1.610099

Cluster 0.
Mobility Characteristics: this cluster has an average mobility score of 37.85, ranging from 26 to 48. This is the highest mobility score of the three clusters.
Demographic Characteristics: the average age of this cluster is 45.62 years, which is the oldest of the three clusters. They have an average education level of 4.42, which is between the other two clusters. Members of this cluster are relatively more engaged in both their professional and private lives, at 65.71 and 68.56, respectively.In addition, they have a mean motivation score of 2.77, which is much higher than the other two clusters.
Cluster 1.
Mobility Characteristics: the average mobility score for this cluster is 29.73, ranging from 13 to 43. This is the lowest mobility score of the three clusters.
Demographic Characteristics: the average age of this cluster is 38.51, which is the youngest of the three clusters. Their average education level is 6.00, which means that the members of this cluster are the most educated. Nonetheless, their participation in both professional and private life is relatively low at 44.60 and 51.57, respectively.Their mean value of motivation is 1.46, which is the lowest among the three clusters.
Cluster 2.
Mobility Characteristics: the average mobility score for this cluster is 33.15, ranging from 20 to 46.

Demographic Characteristics: the average age of this cluster is 42.36 years, which is between the other two clusters. Their average educational attainment is 3.51, which is the lowest among the three clusters. Their involvement in professional and private life is 47.81 and 53.14 respectively, these values are slightly higher than Cluster 1. Their mean value of motivation is 1.61 which is slightly higher than Cluster 1 but much lower than Cluster 0.

So synthesizing all the information, Cluster 0 represents the older, active and motivated population; Cluster 1 represents the younger but more educated, less active and less motivated population; and Cluster 2 represents the middle-aged, less educated but slightly more active and motivated population. These three clusters reveal underlying patterns in the data for us, showing correlations between mobility, age, education, and activity.

Here we can also change the number of other clusters to explore and discover relationships between variables. I then evaluated the number of different clusters with the help of the elbow method and contour scores.



Elbow Method plot.

DESCRIPTION: This graph shows the Inertia values corresponding to different number of clusters (from 2 to 10) Inertia is the sum of the squares of the distances from each point to the center of the cluster to which it is assigned, so lower values of Inertia indicate that the points in the cluster are closer together.

Analysis: From the description, we can see that the Inertia value drops from 60,000 to about 21,000. this is a typical "elbow" shape, where the drop in Inertia begins to slow down, suggesting that adding more clusters may not significantly improve the performance of the model. Finding the "elbow" or inflection point is a common way to determine the optimal number of clusters.

Silhouette Scores plot.

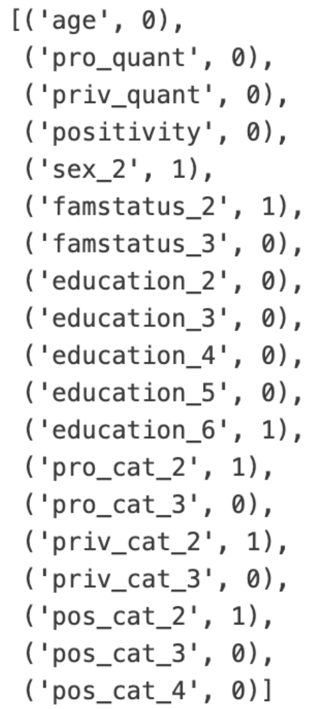
Description: This plot shows the silhouette scores for different numbers of clusters. Silhouette scores range from [-1,1]. High contour scores indicate that points within clusters are closer to each other and further away from points in other clusters.

ANALYSIS: From the description, the contour score reaches a maximum value of 0.345 at 2 clusters and then drops sharply and fluctuates with the increase in the number of clusters. This indicates that the data are best clustered when there are only two clusters. Nonetheless, when the number of clusters is 3, the contour score decreases rapidly, which may mean that 3 clusters is not a good choice.

So, while the Elbow Method plot may imply that there are more than 2 clusters, the contour score plot clearly shows that 2 clusters provide the best aggregation for the data. Dividing the data into 2 clusters may be optimal because it not only results in lower Inertia values, but also in the highest contour scores.
When adding more clusters, the performance may decrease as the internal consistency of the data decreases and the separation from other clusters is not significantly improved. Therefore, for this dataset, selecting 2 clusters for K-means clustering may be the most appropriate.

V. Apply unsupervised learning on the variables

Then I performed unsupervised classification of the feature values, I tried to classify them into 2, 3, and 4 classes and then obtained many different results, here I give a rough description of the classification into 2 clusters.



```
[('age', 0),
 ('pro_quant', 0),
 ('priv_quant', 0),
 ('positivity', 0),
 ('sex_2', 1),
 ('famstatus_2', 1),
 ('famstatus_3', 0),
 ('education_2', 0),
 ('education_3', 0),
 ('education_4', 0),
 ('education_5', 0),
 ('education_6', 1),
 ('pro_cat_2', 1),
 ('pro_cat_3', 0),
 ('priv_cat_2', 1),
 ('priv_cat_3', 0),
 ('pos_cat_2', 1),
 ('pos_cat_3', 0),
 ('pos_cat_4', 0)]
```

Cluster 0 primarily includes age, professional and personal quantitative activity metrics (pro_quant and priv_quant), positivity, and most of the education level and categorical metrics.
Similar to the previous Cluster 1, this cluster appears to be highly correlated with basic demographic information and activity engagement.
Cluster 1 primarily relates to female (sex_2), family status - divorcee (famstatus_2), a high level of education (education_6), professional and personal categorical indicators, and certain positivity categories.
This cluster seems to be more related to people's social and cultural background, especially their gender, family status and certain activeness classifications.

Cluster 0 primarily represents those characteristics that are highly correlated with general demographics and activity engagement.
Cluster 1, on the other hand, is more related to people's social and cultural backgrounds, such as their gender, family status, and certain specific educational and activity classifications.

VI. Conclusion

In this project, I began with an exploratory data analysis of the Flourishing dataset. Through descriptive analysis of the data, I gained a basic understanding of the variables within the dataset. I further explored the correlations between the variables, especially the target variable "Flow".

I applied linear regression and KNN regression models to predict the "Flow" variable, and evaluated them using a variety of evaluation metrics. The results show that although both models have some predictive ability, they suffer from the problem of overfitting. In addition, I used logistic regression and KNN classifiers to predict "Flow" categories based on quartiles and evaluated the performance of these models.

In the unsupervised learning section, I applied the K-means clustering method to cluster observations and variables. By comparing different numbers of clusters, I found that dividing the data into two or three clusters may be the best option.

Finally, I clustered the features and found clusters related to demographics, activity participation, education, family, and cultural background.

Overall, this project provided me with deep insights into the Flourishing dataset and helped me understand the major patterns and trends in the data. By applying various machine learning techniques, I gained a more comprehensive and in-depth understanding of the data.