

# Big Data Intelligence Assignment 1

Gausse Mael DONGMO KENFACK (董沫高斯)  
Student ID: 2024403346

October 2024

## Contents

<b>Introduction</b>	<b>1</b>
<b>1 ANOVA</b>	<b>2</b>
1.1 Visualisation, Hypothesis and Assumptions . . . . .	2
1.2 ANOVA Table, Kruskal-Wallis H test and Conclusions . . . . .	3
<b>2 Regression Problems</b>	<b>4</b>
2.1 Simple Linear Regression . . . . .	4
2.1.1 Linear Regression Results . . . . .	4
2.1.2 Observations . . . . .	4
2.2 Weighted Multivariate Linear Regression . . . . .	5
2.2.1 Weighted Multivariate Linear Regression Results . . . . .	5
2.2.2 Observations . . . . .	5
2.3 Classification . . . . .	6
2.3.1 Binary Classification . . . . .	6
2.3.2 Feature Selection . . . . .	6
2.3.3 Multi-Class Classification . . . . .	6
<b>3 Sampling</b>	<b>7</b>
3.1 Simple Random Sampling . . . . .	7
3.2 Stratified Random Sampling . . . . .	7
3.3 Reservoir Sampling . . . . .	7
3.4 Observations . . . . .	7
<b>Conclusion</b>	<b>8</b>
<b>References</b>	<b>8</b>

## Introduction

In this report, we explore the relationships between group categories and behavioral features using statistical techniques. The analysis is conducted following the specific questions outlined in the homework as a guideline. Additionally, this document is accompanied by a *Jupyter Notebook* file containing the complete code for the analysis.

# 1 ANOVA

In this section, we will conduct a one-way ANOVA to analyze the differences in average age across the five group categories.

## 1.1 Visualisation, Hypothesis and Assumptions

First, we group the data by categories and present some descriptive statistics. We then visualize the data using a box plot, leveraging the Python libraries *Pandas*, *Seaborn* and *Matplotlib*. The results are displayed by the table 1 and the figure 1.

Category	Count	Mean	Std. deviation
House & living	196	30.791534	2.552901
Online Game	484	23.404278	4.923757
Organization and Industry	635	28.545081	3.018979
School Alumni	300	29.618193	5.217371
Stock Market	425	26.255318	5.098256

Table 1: Descriptive statistics on the data of column 7

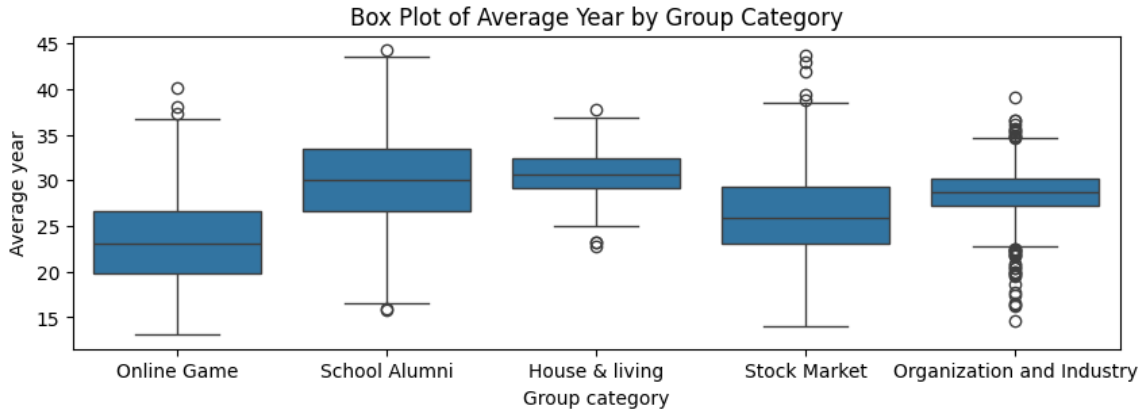


Figure 1: Box Plot of Average Year by Group Category

The hypotheses for our one-way ANOVA in this case are:

- $H_0$  : There is no significant difference in the mean average age between the five categories.
- $H_1$  : At least one category has a mean average age different from the others.

Before conducting our ANOVA analysis, it is crucial to verify that the data meets the necessary assumptions for applying ANOVA. We performed these checks using Python and the *Scipy* library and found that two of the assumptions are not satisfied:

- We applied the Shapiro-Wilk test to assess the normality of the distributions, and the only category out of the five that follows a Gaussian distribution is **House & Living**.

- The descriptive statistics in Table 1 indicate that the categories do not have equal variances, a finding further confirmed by the Levene test.

After obtaining this information, we proceeded with the ANOVA analysis despite the violations, as one-way ANOVA is generally robust to non-Gaussian data. Additionally, we decided to perform a Kruskal-Wallis H test as a non-parametric alternative.

## 1.2 ANOVA Table, Kruskal-Wallis H test and Conclusions

The result obtained for our ANOVA is presented in the Table 2.

Source	SS	df	MS	F	p-value
Between	12782.92	4	3195.73	171.51	1.08e-126
Within	37918.62	2035	18.63		
Total	50701.54	2039			

Table 2: ANOVA Table

The ANOVA results show a significant difference in the means of the categories, as reflected by the large F-value and extremely small p-value. This suggests that the group category has a strong effect on the average age. Since the p-value is lower than 0.05 it suggest to reject the null hypothesis and accept the alternative: **At least one category has a mean average age different from the others**. However, given the earlier violations of normality and homogeneity of variances, it is prudent to also consider the results of the Kruskal-Wallis H test.

The Kruskal-Wallis H test (sometimes also called the "one-way ANOVA on ranks") is a rank-based nonparametric test that can be used in situation like ours when we lack variance homogeneity and normality. We computed the test results with *Scipy* and the results are displayed in Table 3

Statistics	Value
H-statistic	544.10
df	4
p-value	1.93e-116

Table 3: Kruskal-Wallis H test results

The large H-statistic indicates substantial differences in the rank distributions between the groups. The small p-value much less than 0.05 strongly suggests that the null hypothesis can be rejected : **There is at least one different median**.

## 2 Regression Problems

In this section, we will address regression tasks, including simple and weighted multivariate linear regression, as well as classification using logistic regression.

### 2.1 Simple Linear Regression

We began by filtering the data according to the given condition (Session number  $\geq 20$ ) and then conducted a simple linear regression to predict chatting behavior based on the average age of group members. We utilized the *scikit-learn* library to perform the regression.

#### 2.1.1 Linear Regression Results

The results of the Linear Regressions are presented by the table 4.

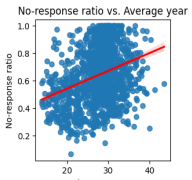
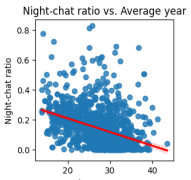
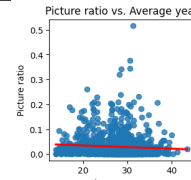
Column	No-response ratio	Night-chat ratio	Picture ratio
Function	$y = 0.0129 * x + 0.29$	$y = -0.0091 * x + 0.39$	$y = -0.0007 * x + 0.05$
RMSE	0.1997	0.1181	0.0460
MSE	0.0399	0.0139	0.0021
Correlation	0.3105	-0.3628	-0.0713
Plot			

Table 4: The 3 Simple Linear Regressions

#### 2.1.2 Observations

The results of the linear regressions reveal some interesting insights into the relationship between group average age and chatting behaviors, along with the prediction errors for each model:

- **No-response ratio** : The correlation of 0.3105 shows a moderate positive relationship between the average age and the no-response ratio, indicating that older people are slightly more likely to not respond. However, the mean squared error of 0.0399 indicate that the model has a relatively moderate prediction error, meaning the predictions are not highly accurate.
- **Night-chat ratio** : The negative correlation of -0.3628 suggests that younger people tend to engage more in night chats, with older people showing less night-chat activity. The mean squared error of 0.0139 is smaller compared to the no-response ratio, indicating a better fit but still not accurate enough for night-chat behavior.
- **Picture ratio** : The weak correlation of -0.0713 implies that age has almost no relationship with picture-sharing behavior. The mean squared error of 0.0021 suggest the model has the lowest prediction errors among the three regressions; this may be due to a lower standard deviation in the data.

## 2.2 Weighted Multivariate Linear Regression

In this problem, we aim to predict the same columns using 8 features (columns 3 to 10), with column 11 serving as weights. Instead of fitting a line, we are now looking for a hyperplane, utilizing the *scikit-learn* library for this task. As visualization becomes challenging with more than 3 dimensions, we won't be plotting the results. Instead, we will focus on discussing the linear functions and the outcome of the analysis.

### 2.2.1 Weighted Multivariate Linear Regression Results

In the results presented here the variable  $x_i$  represent the column  $i + 2$ ; the metrics chosen are the mean squared errors as previously but here we are also interested in how well the regression predictions approximate the data so we calculated the  $R^2score$  of the models.

- **No-response ratio**

$$y = 0.63 + 0.63 * x1 + 0.48 * x2 + 0.67 * x3 + 0.63 * x4 + 0.63 * x5 + 0.61 * x6 + 0.40 * x7 + 0.63 * x8$$

$$RMSE = 0.1439, \quad MSE = 0.0207, \quad R^2score = 0.5218$$

- **Night-chat ratio**

$$y = 0.27 + 0.27 * x1 + 0.32 * x2 + 0.30 * x3 + 0.26 * x4 + 0.28 * x5 + 0.22 * x6 + 0.40 * x7 + 0.27 * x8$$

$$RMSE = 0.0877, \quad MSE = 0.0077, \quad R^2score = 0.3011$$

- **Picture ratio**

$$y = 0.04 + 0.04 * x1 + 0.04 * x2 + 0.06 * x3 + 0.04 * x4 + 0.04 * x5 + 0.03 * x6 + 0.03 * x7 + 0.04 * x8$$

$$RMSE = 0.0464, \quad MSE = 0.0022, \quad R^2score = 0.1037$$

### 2.2.2 Observations

- **No-response ratio** : The coefficients associated with each feature (ranging between 0.48 and 0.67) suggest a strong and relatively uniform contribution from most of the features. The strongest impact comes from features like the Density and the Sex ratio in the group, indicating they may have a more substantial influence on the no-response ratio. For the metrics, the MSE is low enough to say that the model makes less error than the simple regression one. Also, the high  $R^2score$  is a good feature cause it means the model can explain 52% of the variance in the no-response behavior.
- **Night-chat ratio** : For the night-chat ratio, the influence is more varied. It's mostly influenced by the age gap but some other columns are also significant. The MSE is very low so the model makes smaller prediction errors. However the  $R^2score$  is not as high as the one for the no-response ratio.
- **Picture ratio** : In the case of picture ratio, the coefficients are much smaller, all below 0.06, indicating a lower overall influence of the features on this behavior. None of the features appear to strongly affect the picture ratio, which implies that the columns are not really correlated with habit of sharing picture. Similarly to the simple linear regression the MSE is still the lowest, so less errors; unfortunately it doesn't really describe the variance in picture sharing ratio since the  $R^2score$  is only 10%.

## 2.3 Classification

In this task, our goal is to predict the category of a group chat based on other available features. We will begin by approaching this as a binary classification problem, focusing on distinguishing between categories 1 and 4 using logistic regression. After that, we will perform feature selection to identify which features contribute most to improving the prediction accuracy. Finally, we will extend the analysis to a multi-class classification, attempting to classify all categories. 80% of the data are used for the training and 20% for the validation.

### 2.3.1 Binary Classification

After filtering the data to keep only the group chats from category 1 and 4, we performed the binary classification with all the features available and here are the results:

*Precision* : 0.83, *Recall* : 0.81, *F1 – score* : 0.82, *Accuracy* : 0.80

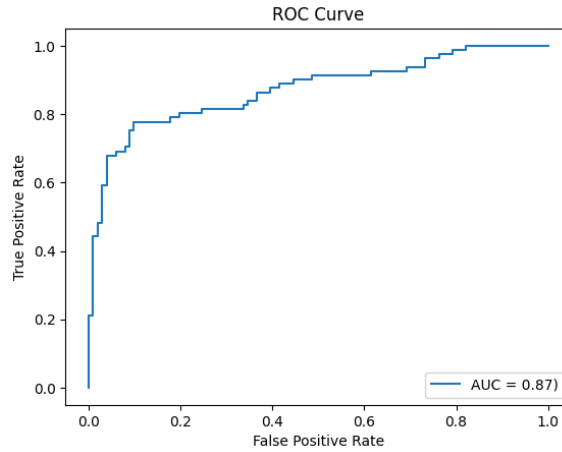


Figure 2: ROC Curve of the logistic regression model

We can notice that the model has pretty good performances on the validation set. Now let's see if we can do better with a wise feature selection.

### 2.3.2 Feature Selection

We performed 12 Recursive Feature Eliminations (RFE) to determine the optimal number of feature and for our data split, to increase the F1-score to the optimum we need **10 features**. The code shows the features selected and also their coefficients and odds ratio. In that case, here are the results:

*Precision* : 0.90, *Recall* : 0.88, *F1 – score* : 0.89, *Accuracy* : 0.88

### 2.3.3 Multi-Class Classification

Finally we used all the data set to perform a multi-class logistic regression to predict all the category. The results are less convincing:

*Precision* : 0.60, *Recall* : 0.62, *F1 – score* : 0.59, *Accuracy* : 0.62

### 3 Sampling

The objective of the final problem is to compare three sampling methods: simple random sampling, stratified random sampling and reservoir sampling. For each method, we will sample 10% of the data, perform 10 linear regressions to predict the no-response ratio based on average age, and then calculate the mean and variance of the regression coefficients  $w_0$  and  $w_1$ .

#### 3.1 Simple Random Sampling

To perform the simple random sampling we used the *sample* function of the *pandas* library which return a random sample of the data. we the performed the regression and got these results:

$$w_0 - \text{mean} : 0.318, \quad w_0 - \text{variance} : 0.0042, \quad w_1 - \text{mean} : 0.0120, \quad w_1 - \text{variance} : 6.08\text{e-}6$$

#### 3.2 Stratified Random Sampling

To perform the stratified random we proceeded similarly as the simple random but first we grouped the data by category; conducting to these results:

$$w_0 - \text{mean} : 0.290, \quad w_0 - \text{variance} : 0.0096, \quad w_1 - \text{mean} : 0.0129, \quad w_1 - \text{variance} : 1.48\text{e-}5$$

#### 3.3 Reservoir Sampling

For the reservoir sampling we wrote the algorithm described in the lecture note.

$$w_0 - \text{mean} : 0.308, \quad w_0 - \text{variance} : 0.0048, \quad w_1 - \text{mean} : 0.0121, \quad w_1 - \text{variance} : 6.65\text{e-}6$$

We also made a box plot of the distribution of the regression parameters for each sampling method.

#### 3.4 Observations

By analyzing the statistics, we can immediately observe that, in this run, Simple Random Sampling emerges as the most stable method, achieving the lowest variance for both  $w_0$  and  $w_1$ . This is closely followed by Reservoir Sampling.

When comparing these results to those in Table 4, it's remarkable that using only 10% of the data yields results very close to those obtained with the full dataset. This demonstrates the efficiency of sampling, where we achieve almost the same level of accuracy with a fraction of the data. Additionally, Stratified Random Sampling performed exceptionally well by predicting values identical to those in Table 4, highlighting its effectiveness in capturing the structure of the data.

These findings emphasize the significance and power of sampling techniques. By reducing the amount of data processed, we not only decrease computation time and resource usage but also retain high-quality predictions. This is particularly important for large datasets, where full regression may be computationally expensive, and smart sampling can offer an efficient alternative.

## Conclusion

In conclusion, this assignment provided a broad exploration of essential concepts in data analysis, including regression, classification, and sampling methods. By working through each problem, we deepened our understanding of how different techniques can be applied to real-world data. It was highly engaging, allowing us to apply theoretical concepts to practical problems. It was not only a great exercise in coding and statistical analysis but also in understanding how to effectively model data and make predictions in real-world scenarios. The hands-on experience with tools like *scikit-learn* reinforced my learning, making this an invaluable exercise.

## References

- Lecture 2 : Sampling and ANOVA
- Lecture 3 : Regression
- <https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php>
- <https://www.geeksforgeeks.org/how-to-perform-a-kruskal-wallis-test-in-python/>
- <https://www.geeksforgeeks.org/understanding-feature-importance-in-logistic-regression-models/>