VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF APPLIED SCIENCE

ANALYSIS ON POPULARITY OF PROGRAMMING LANGUAGES

Supervisor: Ms. Phan Thị Hường
Course: Probability and Statistics (MT2013)
Group: CC03 - Group 06

Ho Chi Minh City, May 11th 2023

TABLE OF CONTRIBUTION

	Members	Student ID	Contribution
1	(Leader) Lê Quang Khải Email: khai.lequang2003@hcmut.edu.vn Tel.: 0905946785	2152659	20%
2	Võ Trần Minh Hiếu	2152560	20%
3	Trương Hải Nam	2150034	20%
4	Dương Trọng Phúc	2152237	20%
5	Hồ Ngọc Bảo Quỳnh	2152935	20%

COMMENT AND EVALUATION

Comment	Evaluation

Contents

I. Data introduction

The dataset used in this report contains information on over 4000 programming languages. It is taken from kaggle.com under the form of a csv file. In this dataset, there are 353 columns that indicate facts about each language such as the year of appearance, year of the latest use, number of users, number of jobs, etc. and 4818 rows for all languages listed in this database.

Table of variables:

No.	Name of variable	Description
1	title	Name of the programming language in a row
2	appeared	The year the programming language was invented
3	type	Type of language
4	lastActivity	The latest year the programming language was used
5	bookCount	The number of book with the topic related to the programming language
6	paperCount	The number of researching papers related to the programming language
7	numberOfUsers	The number of users of the programming language
8	numberOfJobs	The number of jobs that require knowledge of the programming language

II. Background

1) Simple linear regression

The case of simple linear regression considers a single predictor variable or independent variable X and a dependent or response variable Y.

Suppose that for a specified value X of the independent variable the value of the response variable Y can be expressed as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x)

- β_0, β_1 are unknown parameters and called regression coefficients.

- ε is called the random error and assumed to be normally distributed with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$

A simple linear regression model given above states that mean of the random variable Y is related to x by the following straight-line relationship:

$$E(Y|x) = \beta_0 + \beta_1 x$$

where β_0 , β_1 are respectively the intercept and the slope of the straight-line.

2) One-sample t-test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific number. The data used in this test should be a random sample from a normal population with unknown standard deviation. Its test statistic is a Student distribution with n - 1 degrees of freedom (n is the number of objects in the sample):

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

where:

- $\bar{X}$ is the point estimate of the population's average value

- μ is the sample average value

- s is the sample standard deviation

- n is the number of objects in the sample

3) Multiple linear regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. MLR model has the formula:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon$$

where, for i = n observations:

- y_i: dependent variable
- x_i: independent variable
- β_0: y-intercept (constant term)
- β_k: slope coefficients for each independent variable
- ε: the model's error term (also known as residuals)

III. Descriptive statistics

1) Import data

Use read.csv() command to read the file

	title	appeared	type	pldbId	rank	languageRank	factCoun
1	Java	1995	pl	java	0	0	
2	JavaScript	1995	pl	javascript	1	1	
3	C	1972	pl	c	2	2	
4	Python	1991	pl	python	3	3	
5	SQL	1974	queryLanguage	sql	4	4	
6	C++	1985	pl	cpp	5	6	
7	HTML	1991	textMarkup	html	6	5	
8	Linux	1991	os	linux	7	NA	
9	XML	1996	dataNotation	xml	8	7	

2) Data cleaning

- Key variables: title, appeared, type, lastActivity, bookCount, paperCount, numberOfUsers, numberOfJobs. After choosing key variables, we save the database to a variable called “my_data”.

	title	appeared	type	lastActivity	bookCount	paperCount	numberOfUsers	numberOfJobs
1	Java	1995	pl	2022	401	37	5550123	85206
2	JavaScript	1995	pl	2022	351	48	5962666	63993
3	C	1972	pl	2022	78	19	3793768	59919
4	Python	1991	pl	2022	342	52	2818037	46976
5	SQL	1974	queryLanguage	2022	182	37	7179119	219617
6	C++	1985	pl	2022	128	6	4128238	61098
7	HTML	1991	textMarkup	2022	116	7	5570873	69531
8	Linux	1991	os	2018	2	0	3229009	32007
9	XML	1996	dataNotation	2022	151	37	1917452	42277
10	PHP	1995	pl	2022	274	26	2356101	30349
11	Perl	1987	pl	2022	276	9	491984	13482
12	MATLAB	1984	pl	2022	177	35	2661579	32228
13	Ruby	1995	pl	2022	65	13	357730	11438

- Remove irrelevant and redundant observations: Filtering out observations that are not “pl” type and remove duplicated observations in title.

	title	appeared	type	lastActivity	bookCount	paperCount	numberOfUsers	numberOfJobs
1	Java	1995	pl	2022	401	37	5550123	85206
2	JavaScript	1995	pl	2022	351	48	5962666	63993
3	C	1972	pl	2022	78	19	3793768	59919
4	Python	1991	pl	2022	342	52	2818037	46976
6	C++	1985	pl	2022	128	6	4128238	61098
10	PHP	1995	pl	2022	274	26	2356101	30349
11	Perl	1987	pl	2022	276	9	491984	13482
12	MATLAB	1984	pl	2022	177	35	2661579	32228
13	Ruby	1995	pl	2022	65	13	357730	11438
14	C#	2000	pl	2022	3	0	217261	19747
17	Fortran	1957	pl	2022	321	37	165151	1931
19	R	1993	pl	2022	40	9	1075613	14173
22	Go	2009	pl	2022	5	26	525179	6403

- Check missing data by using is.na() function:

```
> na_loc <- apply(is.na(pl_data), 2, which)
> na_loc
integer(0)
```

The result depicted above shows that there are no missing values in key variables. Also, there are no categorical data in the data set.

3) Data visualization

- Summary:

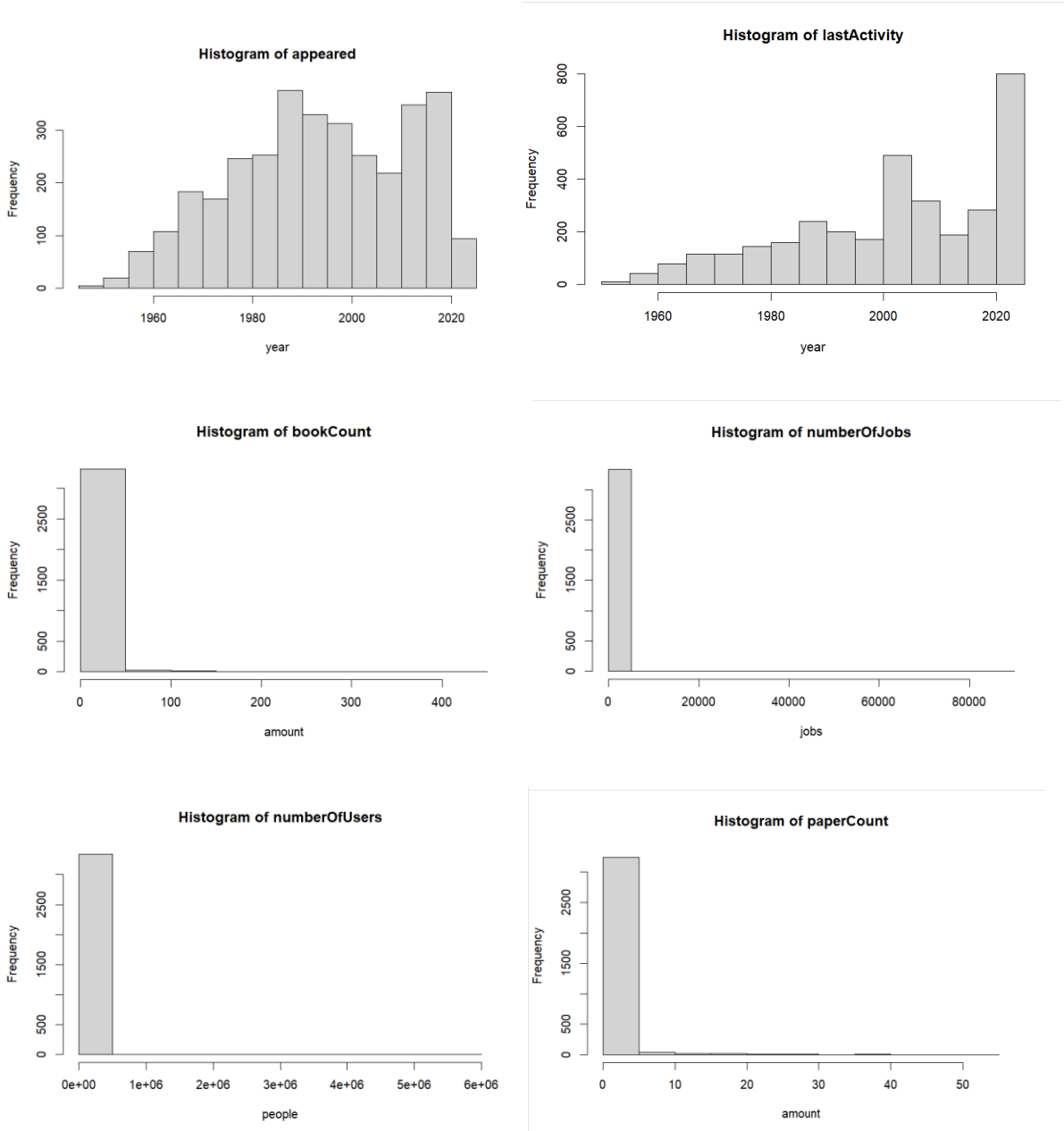
```
   title              appeared        type              lastActivity    bookCount       
 Length:3351        Min.   :1948   Length:3351        Min.   :1951   Min.   :  0.000  
 Class :character   1st Qu.:1982   Class :character   1st Qu.:1989   1st Qu.:  0.000  
 Mode  :character   Median :1994   Mode  :character   Median :2005   Median :  0.000  
                    Mean   :1994                      Mean   :2002   Mean   :  2.127  
                    3rd Qu.:2010                      3rd Qu.:2019   3rd Qu.:  0.000  
                    Max.   :2022                      Max.   :2023   Max.   :401.000  
  paperCount      numberOfUsers       numberOfJobs    
 Min.   : 0.0000   Min.   :      0   Min.   :    0  
 1st Qu.: 0.0000   1st Qu.:      0   1st Qu.:    0  
 Median : 0.0000   Median :     15   Median :    0  
 Mean   : 0.6873   Mean   :  10848   Mean   :  161  
 3rd Qu.: 0.0000   3rd Qu.:    169   3rd Qu.:    0  
 Max.   :52.0000   Max.   :5962666   Max.   :85206  
```

-

Calculate variance and standard deviation of each variable by using var() and sd() function:

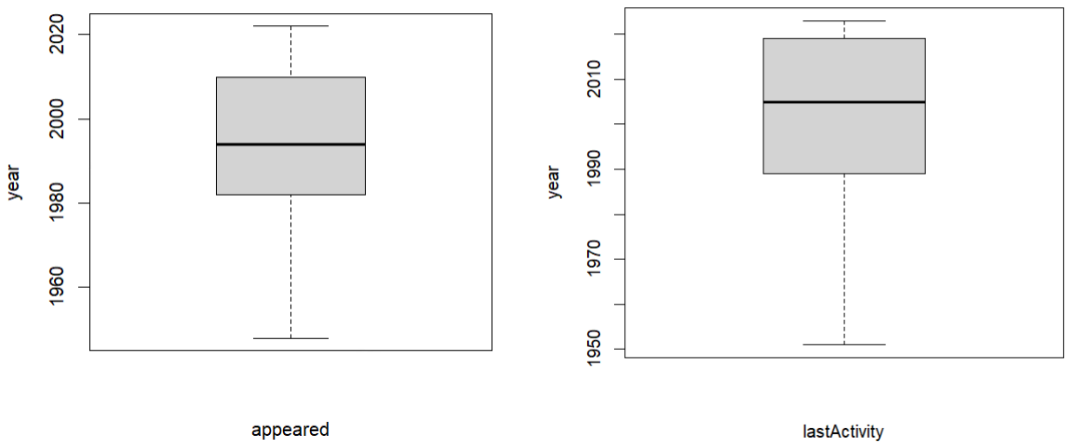
Variable	Variance	Standard deviation
appeared	301.5012	17.36379
lastActivity	320.6715	17.9073
bookCount	341.239	18.47265
paperCount	14.63828	3.826001
numberOfUsers	36357953212	190677.6
numberOfJobs	7287006	2699.445

- Histogram: The histogram is a visual display of the frequency distribution. The histogram provides a visual impression of the shape of the distribution of the measurements and information about the central tendency and scatter or dispersion of the data.

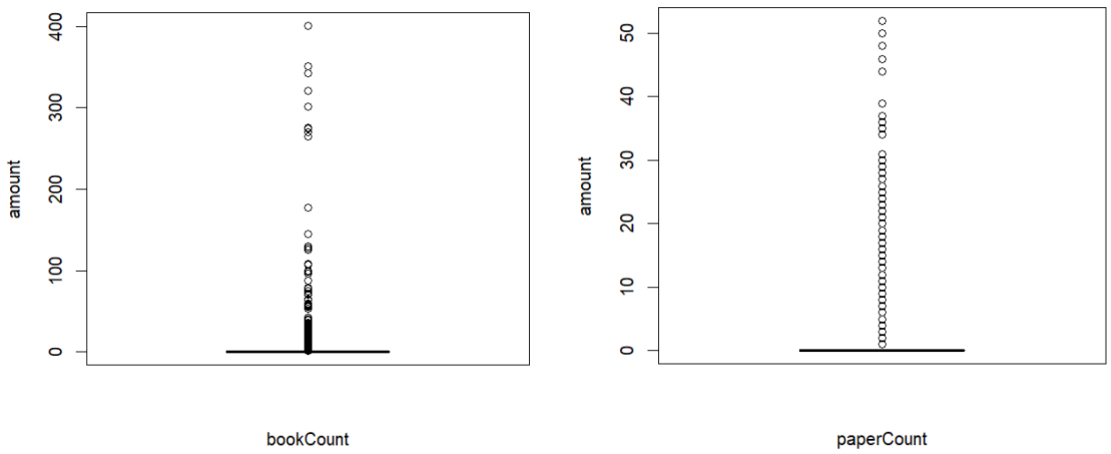


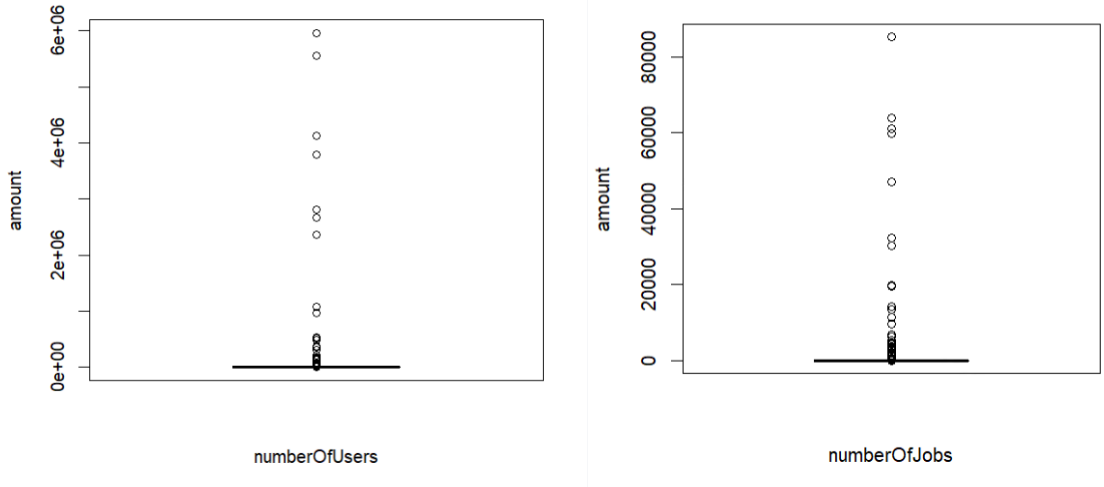
The histogram of appeared and histogram of lastActivity are left skew. The histogram of bookCount, the histogram of paperCount, the histogram of numberOfUsers and histogram of numberOfJobs are right skew.

- Box plot: We construct box plots to examine important features of the dataset, such as center, spread, departure from symmetry, and identification of unusual observations or outliers.



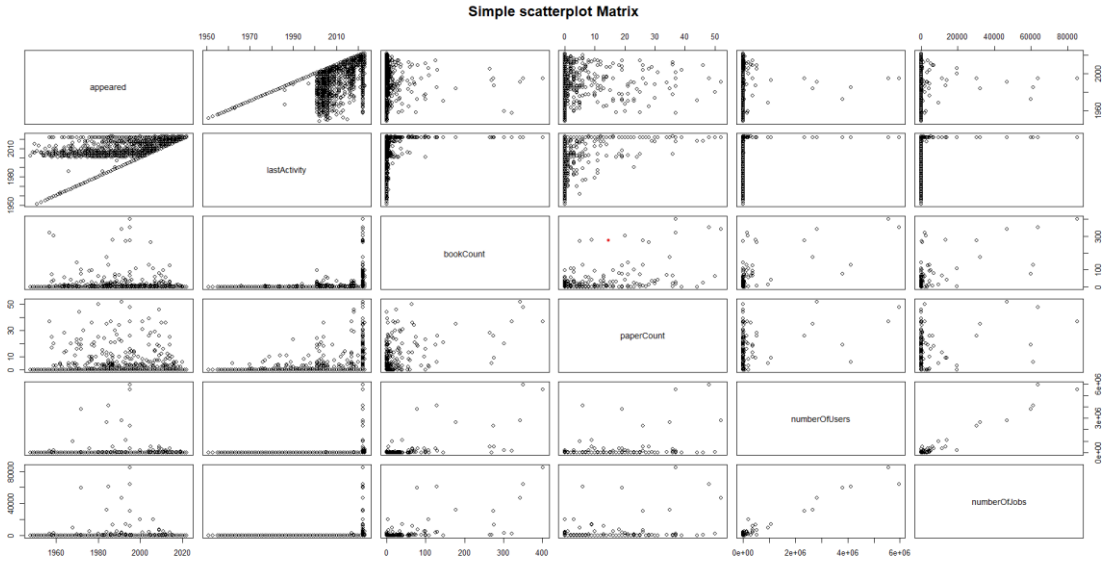
The two first box plots indicate that the time programming languages appeared and the last activity of these programming languages is leaned to the top side.





There are four box plots that the box cannot be seen. This situation happens because the major of the observations in these variables have value 0, the others have large value exceed the limit so the outliers vary extremely to the top.

- Scatter diagram:



According to the matrix of scatter diagrams, there may exist linear relationship between numberOfUsers and numOfJobs, appeared and lastActivity.

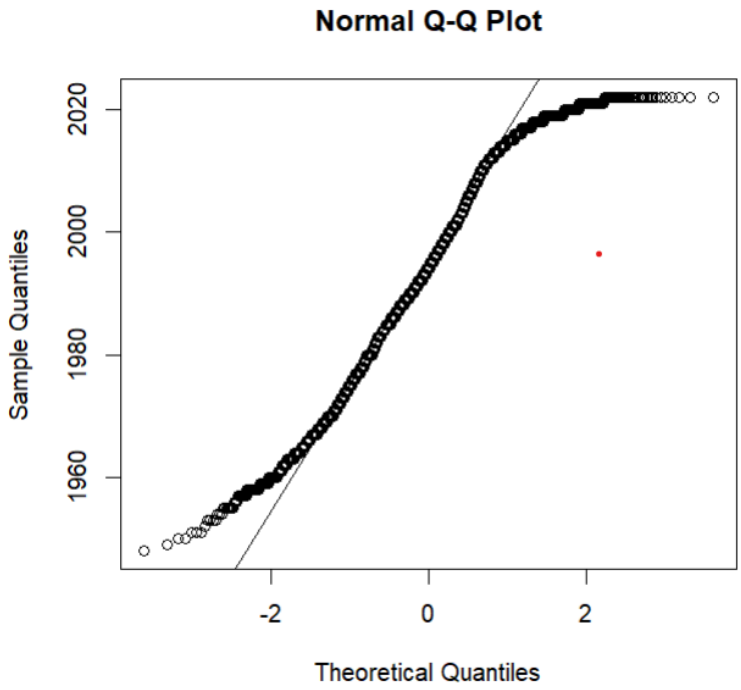
- Sample correlation coefficients of each pair:

	appeared	lastActivity	bookCount	paperCount	numberOfUsers
lastActivity	0.7639				
bookCount	-0.0352	0.1122			
paperCount	-0.0396	0.1393	0.5731		
numberOfUsers	-0.0121	0.063	0.6621	0.384	
numberOfJobs	-0.0138	0.0664	0.6659	0.4016	0.9709

According to the table, bookCount has weak linear relationship with paperCount, numberOfUsers and numberOfJobs. Moreover, there are weak linear relationship between paperCount with numberOfUsers and paperCount with numberOfJobs. However, there exists strong positive linear relationship between numberOfUsers and numberOfJobs. The value of the sample correlation coefficient between appeared and lastActivity is 0.7639. This is a moderately strong correlation, indicating a possible linear relationship between the two variables. The remaining pairs have no relationship.

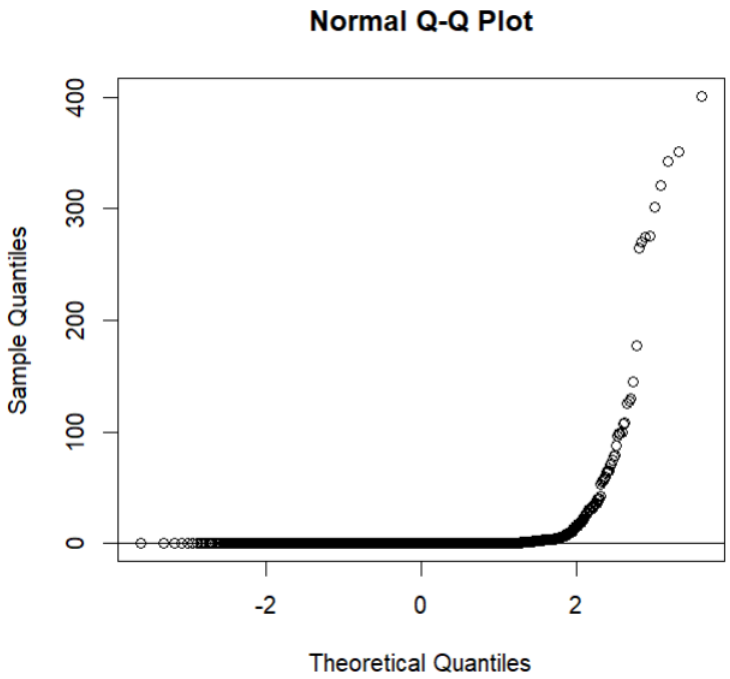
- Probability plot:

+ Appeared:

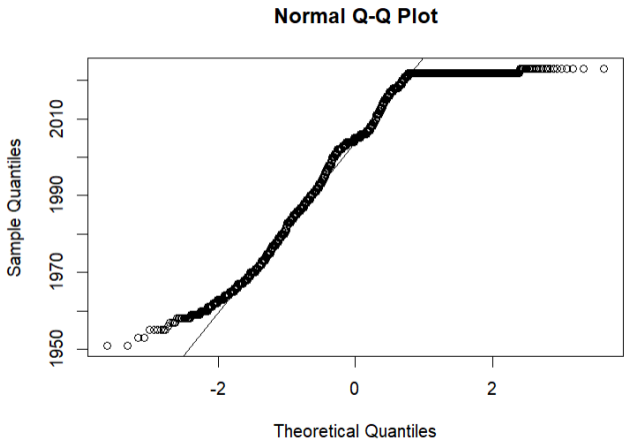


According to the plot, the “appeared” variable follows a light-tailed distribution.

+ bookCount:

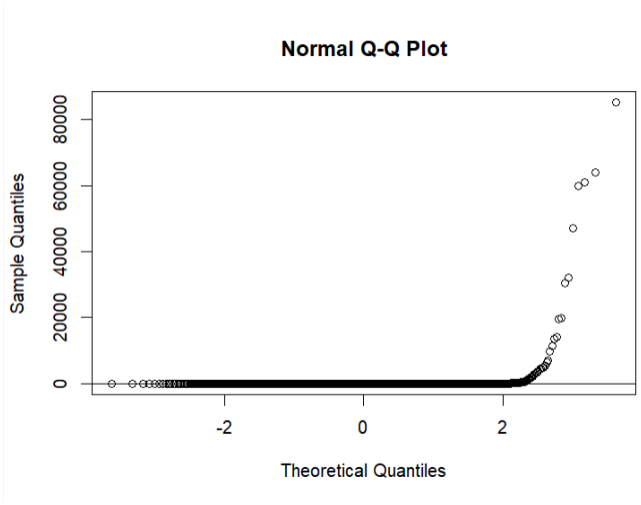


+ lastActivity:

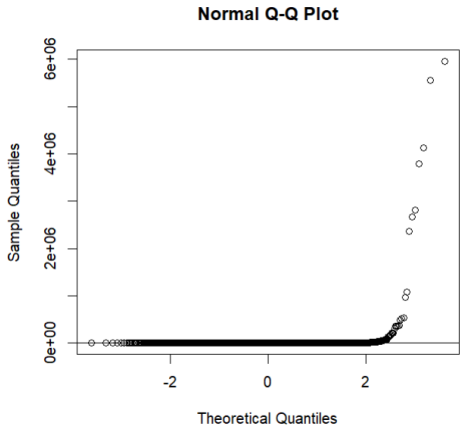


According to the plot, observations in “lastActivity” variable follow a light-tailed distribution.

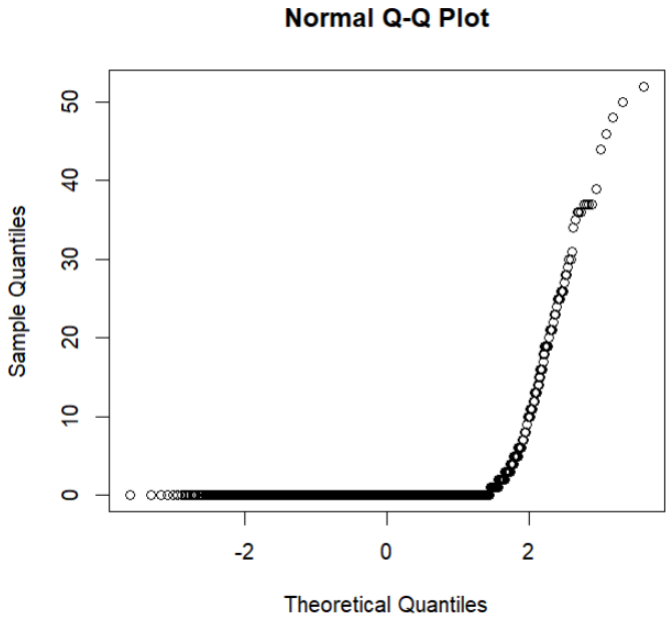
+ numberOfJobs:



+ numberOfUsers:



+ paperCount:



IV. Inferential statistics

1) Is the average number of books equal to 25 or not?

Null hypothesis: $H_0: \mu = 25$: Number of books is equal to 25.

Alternative hypothesis: $H_0: \mu \neq 25$: Number of bookCount is equal to 25.

Testing standard: $\rho - value < 2.2e - 16$

Rejection domain: $W_\alpha = (0; +\alpha); (0;\ 0.05)$

→ $\rho - value \in W_\alpha$

→ In conclusion, we reject H_0

→ The average number of books is not equal to 25.

```
        One Sample t-test

data:  pl_data$bookCount
t = -71.676, df = 3350, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 1.501752 2.753097
sample estimates:
mean of x
 2.127425
```

2) Which factors affect the numberOfJobs most?

```
Call:
lm(formula = numberOfJobs ~ appeared + lastActivity + bookCount +
    paperCount + numberOfUsers, data = pl_data)

Residuals:
     Min       1Q   Median       3Q      Max
-17739.0    -33.2     -0.6      6.5  16747.0

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.517e+02  1.340e+03  -0.337   0.7362
appeared      -2.201e+00  1.021e+00  -2.156   0.0311 *
lastActivity   2.425e+00  9.993e-01   2.426   0.0153 *
bookCount      5.628e+00  9.040e-01   6.226 5.38e-10 ***
paperCount    -1.691e+01  3.596e+00  -4.702 2.68e-06 ***
numberOfUsers  1.350e-02  7.810e-05 172.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 641.8 on 3345 degrees of freedom
Multiple R-squared:  0.9436,    Adjusted R-squared:  0.9435
F-statistic: 1.118e+04 on 5 and 3345 DF,  p-value: < 2.2e-16
```

→ From the table of result above, we can see that correspoding to the estimate column are the estimated coeffients for the regression coefficients. The intercept lines are appeared, lastActivity, bookCount, paperCount, numberOfUsers, respectively.

→Since the p-value for the F-statistic is < 2.2e - 16, at least one of the indenpent variables is significantly related to dependent variable.

→ With 5 variables in total, due to the results of the analysis, it can be seen that changes in lastActivity, bookCount and paperCount have a more significant effect on numberOfJobs than changes in appeared and numberOfUsers.

→ In conclusion, 3 variables: lastActivity, bookCount and paperCount have the most significant effect on numberOfJobs.

3) Find the linearity relationship between numberOfUsers and other variables.

Model of numberOfUsers:

$$y = \beta_0 + \beta_1.appeared + \beta_2.lastActivity + \beta_3.bookCount + \beta_4.paperCount + \beta_5.numberOfJobs$$

```
Call:
lm(formula = numberOfUsers ~ appeared + lastActivity + bookCount +
    paperCount + numberOfJobs, data = pl_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1128195     -487      -27     2568  1536212

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  30040.5946 94148.2342   0.319  0.74969
appeared       210.8532    71.6381   2.943  0.00327 **
lastActivity  -225.7403    70.1387  -3.218  0.00130 **
bookCount      303.2389    63.6477   4.764 1.98e-06 ***
paperCount    1318.0076   252.3702   5.223 1.87e-07 ***
numberOfJobs    66.6090     0.3853 172.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45080 on 3345 degrees of freedom
Multiple R-squared:  0.9442,    Adjusted R-squared:  0.9441
F-statistic: 1.132e+04 on 5 and 3345 DF,  p-value: < 2.2e-16
```

From the result above, we have:

$$\beta_0 = 30040.5946, \beta_1 = 210.8532, \beta_2 = -225.7403, \beta_3 = 303.2389,$$

$$\beta_4 = 1318.0076, \beta_5 = 66.6090$$

So, the model formula can be written as follow:

$$y = 30040.5946 + 210.8532\,.\,appeared - 225.7403\,.\,lastActivity + 303.2389\,.\,bookCount + 1318.0076\,.\,paperCount + 66.6090\,.\,numberOfJobs$$

Hypothesis test

+ Null hypothesis: $H_0: \beta_i = 0$: The regression coefficients β_i does not affect numberOfUsers
+ Alternative hypothesis: $H_1: \beta_i \neq 0$: The regression coefficients β_i affects numberOfUsers

→ Pr(> |t|) of these variables: appeared, lastActivity, bookCount, paperCount and numberOfJobs is less than the significance level $\alpha = 0.05$

→ Reject null hypothesis H_0, accept alternative hypothesis: H_1

→ In conclusion, we can see the linearity relationship between numberOfUsers and appeared, lastActivity, bookCount, paperCount and numberOfJobs through the following equation:

$$y = 30040.5946 + 210.8532\,.\,appeared - 225.7403\,.\,lastActivity + 303.2389\,.\,bookCount + 1318.0076\,.\,paperCount + 66.6090\,.\,numberOfJobs$$

4) Summary

First, we use one pair t-test to check if the average number of books is equal to 25 or not and we get the alternative hypothesis.

Then, by using multiple linear regression model, we are able to determine 3 main independent variables which have the most significant effect on the dependent variable numberOfJobs, they are lastActivity, bookCount and paperCount.

Lastly, we continue using multiple linear regression to find out the linearity relationship between numberOfUsers and other independent variables. As a result, we have all the coefficents corresponding to the variables and get a complete equation.

V. Discussion and Extension

1) Discussion

1.1. T-test

Independent sample t-test is used to compare two sample means from unrelated groups without knowing the standard deviation. However, it can only be used when there are two groups of data, no more or less so when you want to use it on a group of three or more, you have to divide them into group of 2 and then compare it respectively, which is time-consuming and inconvenient. Besides that, the correction of t-test is optimized when the sample size is small (less than 30). In terms of large size sample, z-test (in case you know the standard deviation) and ANOVA (for group of many samples)

1.2. Linear regression

In the previous section we applied multiple linear regression to analyze the linearity relationship between numberOfUsers and the other independent variables. The R-squared of the model is 0.9442 which indicates that the relationship between the dependent and independent variables is somehow linear. So linear regression is an acceptable method to use. The multiple linear regression has some certain benefits like it perform exceptionally well for linearly separable data, it is easy to implement and to interpet the output coefficients. Compared to other algorthims, mulitple linear regression is one of the best to use due to its simplicity. On the other hand, it still has some drawbacks which may affect the result like it is quite often prone to noise and overfitting, quite sensitive to outliers or prone to multicollinearity. In this case, the result we may not get high accuracy or the linear regression model is not fitted with the dataset.

2) Extension

In this section, we will apply a machine learning model which can overcome the drawbacks of linear regression. Polynomial provides the best approximation of the relationship between dependent and independent variables. We will use both linear regression and polynomial regression to visualize the relationship between numberOfJobs and other independent variables and then compare the result.

First we import all the nescessary libraries:

```
#import the libraries
import sys
import matplotlib
import pandas as pd
import numpy as np
matplotlib.use('Agg')

import matplotlib.pyplot as plt
from skLearn.model_selection import train_test_spit
from sk.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
```

Then we read the data, dividing it into 2 components

```
#Import dataset
datas = pd.read_csv('"C:\Users\DELL\Downloads\BTL.csv"')

#Divide data into 2 components
jobs = ['numberOfJobs']
other = ['appeared', 'lastActivity', 'bookCount', 'paperCount',
'numberOfUsers']
X = datas.iloc[:, other].values
y = datas.iloc[:, jobs].values
```

Fitting Linear Regression and Polynomial Regression to the dataset

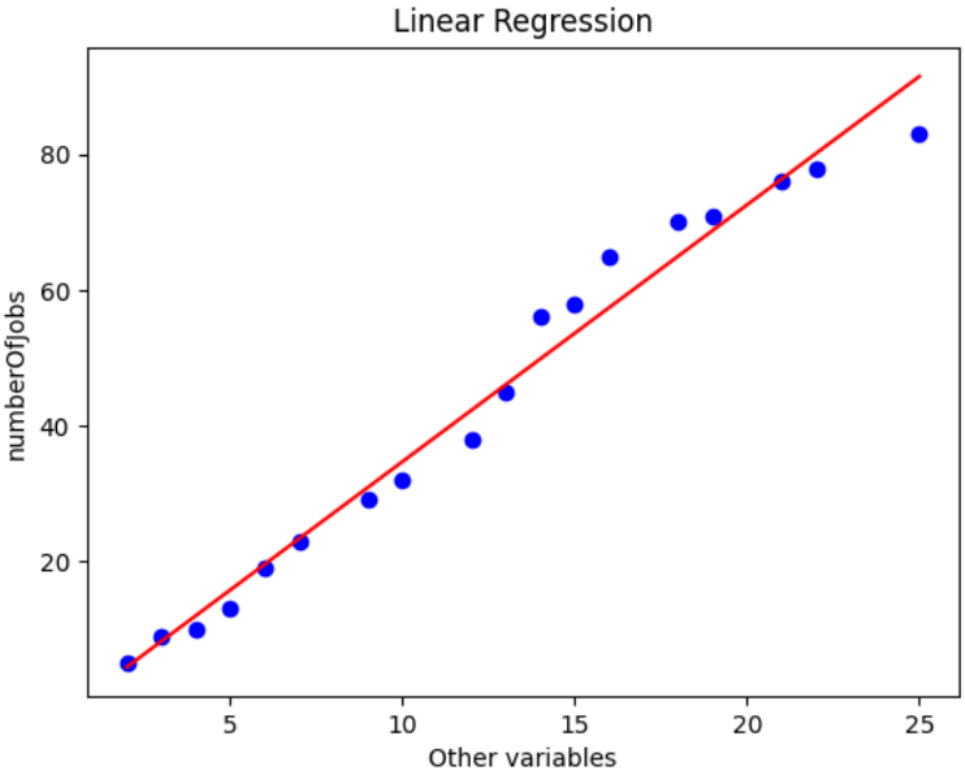
```
#Fitting Linear Regression to the dataset
lin = LinearRegression()
lin.fit(X, y)

#Fitting Polynomial Regression to the dataset
poly = PolynomialFeatures(degree = 4)
X_poly = poly.fit_transform(X)

poly.fit(X_poly, y)
lin2 = LinearRegression()
lin2.fit(X_poly, y)
```

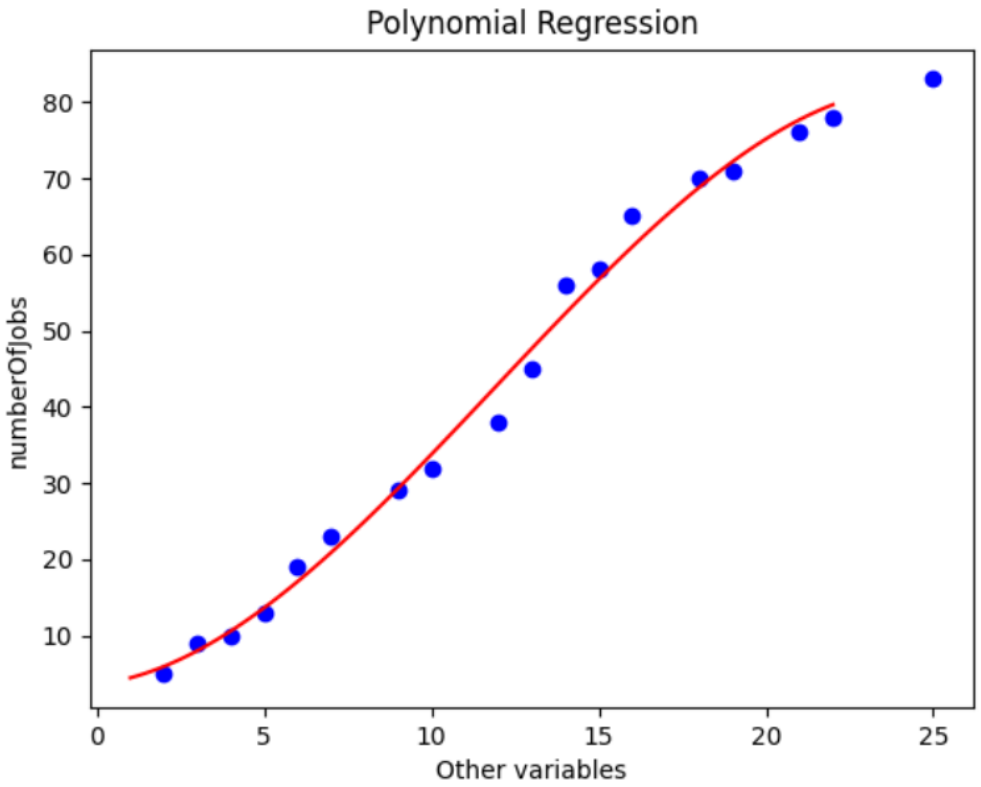
Visualizing the Linear Regression results

```
# Visualising the Linear Regression results
plt.scatter(X, y, color = 'blue')
plt.plot(X, lin.predict(X), color = 'red')
plt.title('Linear Regression')
plt.xlabel('Other variables')
plt.ylabel('numberOfJobs')
plt.show()
```



Visualizing the Polynomial Regression results

```
# Visualising the Polynomial Regression results
plt.scatter(X, y, color = 'blue')
plt.plot(X, lin2.predict(poly.fit_transform(X)), color = 'red')
plt.title('Polynomial Regression')
plt.xlabel('Other variables')
plt.ylabel('numberOfJobs')
plt.show()
```



As you can see from the result, Polynomial Regression give us a more accurate model. The Polynomial Regression model provides the best approximation of the relationship between dependent and independent variables and is the most suitable model for this dataset.

VI. Code and data availability

- Data:

https://www.kaggle.com/datasets/sujaykapadnis/programming-language-database

- Code:

https://drive.google.com/file/d/1qZ2AZMV7O8WhoIZ5Tm54EV9-W1Y5DjEw/view?usp=sharing

VII. References

[1] Douglas C. Montgomery, and George C. Runger. (2014). *Applied Statistics and Probability for Engineers, 6th edition*. Wiley

[2] Peter Dalgaard. (2008). *Introductory Statistics with R, 2nd edition*. Springer

[3] Datacamp. (2022). *Linear Regression in R Tutorial*. Access from https://www.datacamp.com/tutorial/linear-regression-R

[4] Nguyễn Tuyết Anh. (2023). *Kiểm định T - Test | Kiểm định sự khác biệt trong SPSS*. Access from https://luanvan1080.com/kiem-dinh-t-test-trong-spss.html

[5] Phạm Lộc Blog. *Kiểm định Independent Sample T-Test trong SPSS*. Access from https://www.phamlocblog.com/2017/07/kiem-dinh-independent-sample-t-test-SPSS.html

[6] CueMath. *Inferential Statistics. Access from* https://www.cuemath.com/data/inferential-statistics/

[7] Andriy Blokhin. (2022). *Linear vs. Multiple Regression: What's the Difference?*. Access from https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp