ALY 6010 STUDENT SURVEY DATA ANALYSIS

ALY 6010 : Probability Theory and Introductory Statistics

WEEK 2 R-PRACTICE

Student Name: Godliver Alangyam Awonlie

NUID: 002807730

Date: 11/10/2024

Part 1 :

The dataset is a survey collected from students enrolled in ALY 6010. It includes information on students' undergraduate programs of study and their proficiency levels in software tools such as Python, Excel, and R for data analysis.

I imported the CSV file into R and reviewed the dataset to understand the data types, number of rows, and columns. The dataset consists of 14 rows and 16 columns.

```
Rows: 14 Columns: 16
── Column specification ──────────────────────────────────────────────────────
Delimiter: ","
chr (7): section, section_sis_id, submitted, 9597661: What is your undergraduate degree major?, 959766...
dbl (9): section_id, attempt, 1...7, 1...9, 1...11, 1...13, n correct, n incorrect, score
```

After examining the dataset, I realized the need to restructure its columns and rows. I then created a new dataframe containing the most relevant information.

```
> glimpse(Student_survey)
Rows: 14
Columns: 16
$ section                                                                          <chr> …
$ section_id                                                                       <dbl> …
$ section_sis_id                                                                   <chr> …
$ submitted                                                                        <chr> …
$ attempt                                                                          <dbl> …
$ `9597661: What is your undergraduate degree major?`                              <chr> …
$ `1...7`                                                                          <dbl> …
$ `9597662: How familiar are you with Excel? Use the scale 1 to 5 (with 5 being the expert level)` <chr> …
$ `1...9`                                                                          <dbl> …
$ `9597663: How familiar are you with R? Use the scale 1 to 5 (with 5 being the expert level)?`   <chr> …
```

Part 2: Discriptive Statistics[1][2][3]

After restructuring the dataset, I added 5 new columns with 14 rows each and performed descriptive statistics on the 3 major skillsets in the dataset. The results show that the mean skill level for Excel is 2.78, the highest among the three, followed by R at 2.29, and Python at 1.50. This suggests that most students have limited skills or knowledge in the Python programming language.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Excel_Skill	1	14	2.785714	0.8925824	3	2.833333	1.4826	1	4	3	-0.2213883	-0.9487799	0.2385527
R_Skill	2	14	2.285714	0.4688072	2	2.250000	0.0000	2	3	1	0.8488760	-1.3617347	0.1252940
Python_Skill	3	14	1.500000	0.8548504	1	1.333333	0.0000	1	4	3	1.7151134	2.3334982	0.2284684

I then performed descriptive statistics for the various majors. The data is grouped into four main categories based on students' undergraduate majors: Major 1 represents Biology/Science, Major 2 represents Business, Major 3 represents Humanities, and Major 4 represents STEM. From the analysis, it appears that students with a STEM background have a higher mean skill level across all three software tools compared to those from non-STEM majors, indicating greater proficiency and knowledge.

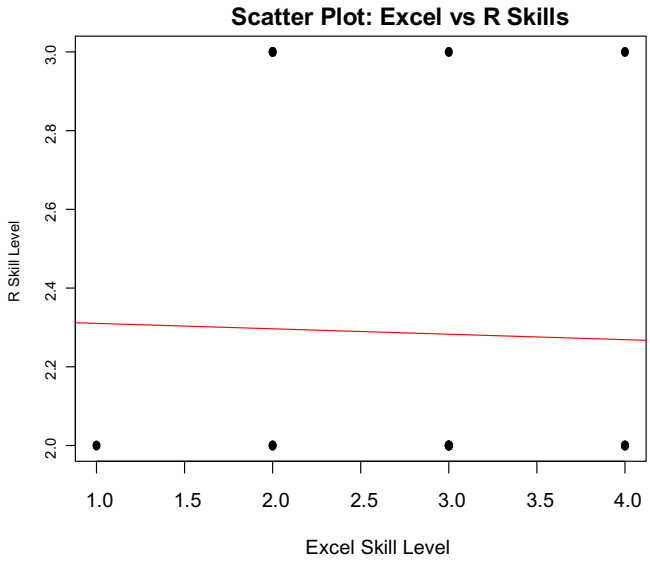
```
> print(group_stats)
df$Major_group: 1
             vars n mean   sd median trimmed mad min max range skew kurtosis  se
Excel_Skill     1 5  3.2 0.45      3     3.2   0   3   4     1 1.07    -0.92 0.2
R_Skill         2 5  2.0 0.00      2     2.0   0   2   2     0  NaN      NaN 0.0
Python_Skill    3 5  1.2 0.45      1     1.2   0   1   2     1 1.07    -0.92 0.2
------------------------------------------------------------------------------------------------------------------------
df$Major_group: 2
             vars n mean   sd median trimmed mad min max range  skew kurtosis   se
Excel_Skill     1 3 2.33 0.58      2    2.33   0   2   3     1  0.38    -2.33 0.33
R_Skill         2 3 2.67 0.58      3    2.67   0   2   3     1 -0.38    -2.33 0.33
Python_Skill    3 3 1.67 0.58      2    1.67   0   1   2     1 -0.38    -2.33 0.33
------------------------------------------------------------------------------------------------------------------------
df$Major_group: 3
             vars n mean sd median trimmed  mad min max range skew kurtosis   se
Excel_Skill     1 3    2  1      2       2 1.48   1   3     2    0    -2.33 0.58
R_Skill         2 3    2  0      2       2 0.00   2   2     0  NaN      NaN 0.00
Python_Skill    3 3    1  0      1       1 0.00   1   1     0  NaN      NaN 0.00
------------------------------------------------------------------------------------------------------------------------
df$Major_group: 4
             vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
Excel_Skill     1 3 3.33 1.15      4    3.33 0.00   2   4     2 -0.38    -2.33 0.67
R_Skill         2 3 2.67 0.58      3    2.67 0.00   2   3     1 -0.38    -2.33 0.33
Python_Skill    3 3 2.33 1.53      2    2.33 1.48   1   4     3  0.21    -2.33 0.88
```

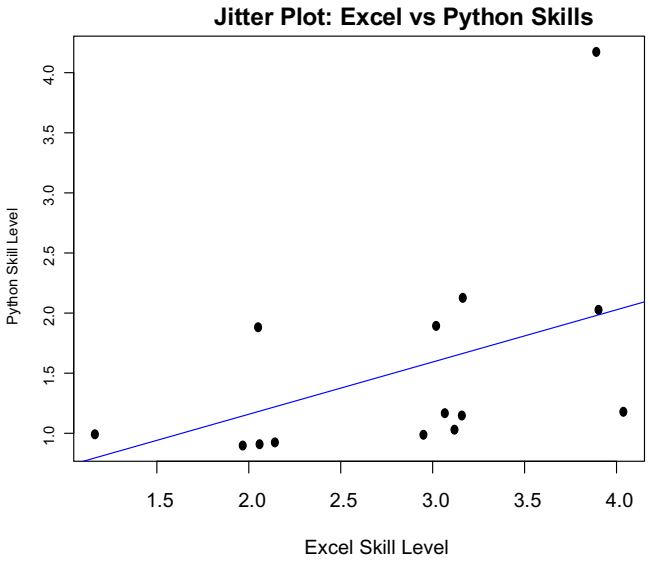
A more simplified discriptive statistics of the data.Excel has the highest mean with a high standard deviation

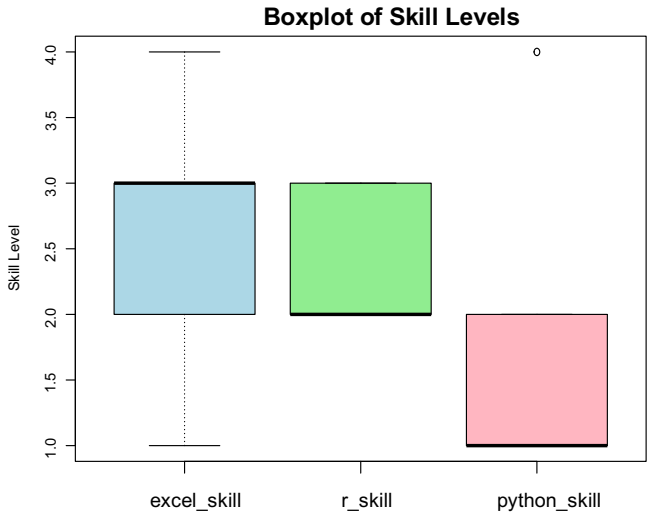
	Statistic	Excel	R	Python
1	Mean (SD)	2.79 (0.89)	2.29 (0.47)	1.50 (0.85)
2	Range	1–4	2–3	1–4
3	N	14	14	14

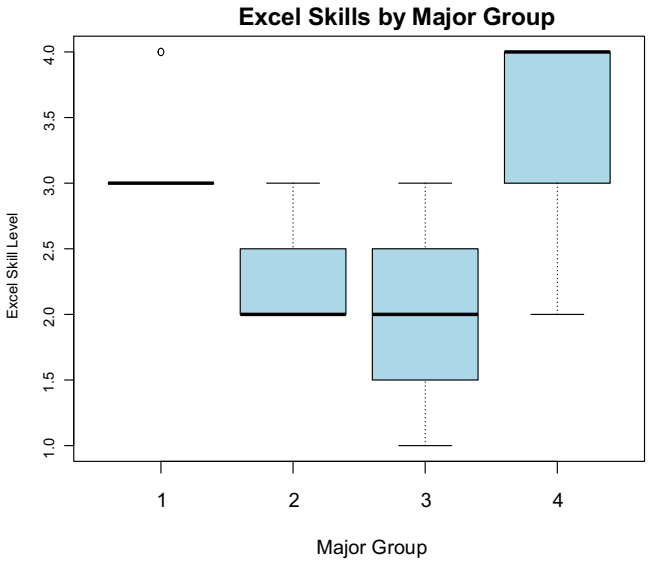
Part 3: Visualisations [4][5][6][7]

The following visualizations shows the relationship between the various skillsets.









Scatter Plot: The scatter plot shows a weak positive correlation between Excel and R skills. The relationship suggests that students with higher Excel proficiency tend to have slightly higher R skills**.**

Jitter Plot: The jitter plot is usually used when dealing with discrete data points that overlap. The jitter adds random noise to help visualize the density of observations particularly useful for this dataset with whose skill levels are discrete.Shows the clustering of Python skills at lower levels regardless of Excel proficiency.

Boxplots: The overall skill distribution of the boxplot shows Excel has the largest spread while R skills are more concentrated with Python skills showing as positive skewness and several outliers detected with the Python skills.

The Python skill with a level of 4 was identified as an outlier while the excell skills oulier include the “never used” response.The R skill showed no praticular outlier suggesting a more consistent skill level.

Summary:

The student survey highlights the varied technology skill levels among the diverse group of ALY 6010 graduate students, with Excel emerging as the strongest area and Python programming as the weakest. The findings suggest areas for potential skills development to better prepare students for data-driven academic and professional pursuits.

References:

[1]Dataset:Redirecting.(n.d.b).https://northeastern.instructure.com/courses/198023/discussion_topics/2572980

[2]Statistics Globe. (2020, April 6). jitter Function in R (Example) | Add Random Noise toNumericValues|DrawPlotwithJitteredPoints[Video].YouTube.https://www.youtube.com/watch?v=HD3N8-a3-vk

[3] Sisi Wang, Feiping Nie, Zheng Wang, Rong Wang, Xuelong Li, Outliers Robust Unsupervised Feature Selection for Structured Sparse Subspace, IEEE Transactions on Knowledge and Data Engineering, 10.1109/TKDE.2023.3297226, 36, 3, (1234-1248), (2024).

[4]I Kabacoff, R. (2022). Introduction to R. In R in Action, Data Analysis and Graphics with R and Tidyverse (3rd ed., p. 462). Manning Publications Co.

[5]Redirecting.(n.d.c).https://m365.cloud.microsoft/chat?fromcode=bingchat&redirectid=B8EBDAF8B0C4FEA9F9A23821C171DD7&auth=2

[6]Redirecting.(n.d.).https://copilot.cloud.microsoft/?fromcode=bingchat&redirectid=8F4080CC01D54E15AED9F4AE946FB94F&auth=2