**Introduction**

The quality control manager of a manufacturing company has randomly selected 50 of the company's recent projects. Each project has four attributes:

1. PIP number that identifies each project uniquely,
2. A Quality score,
3. A Speed, which is the time duration in days during which the project was completed, and
4. The cost of the project.

The manager then assigns an overall score (from zero to 7) to each project depending on how many of the three criteria have been satisfied by that project. Based on the information above, analyze these 50 projects using probability knowledge.

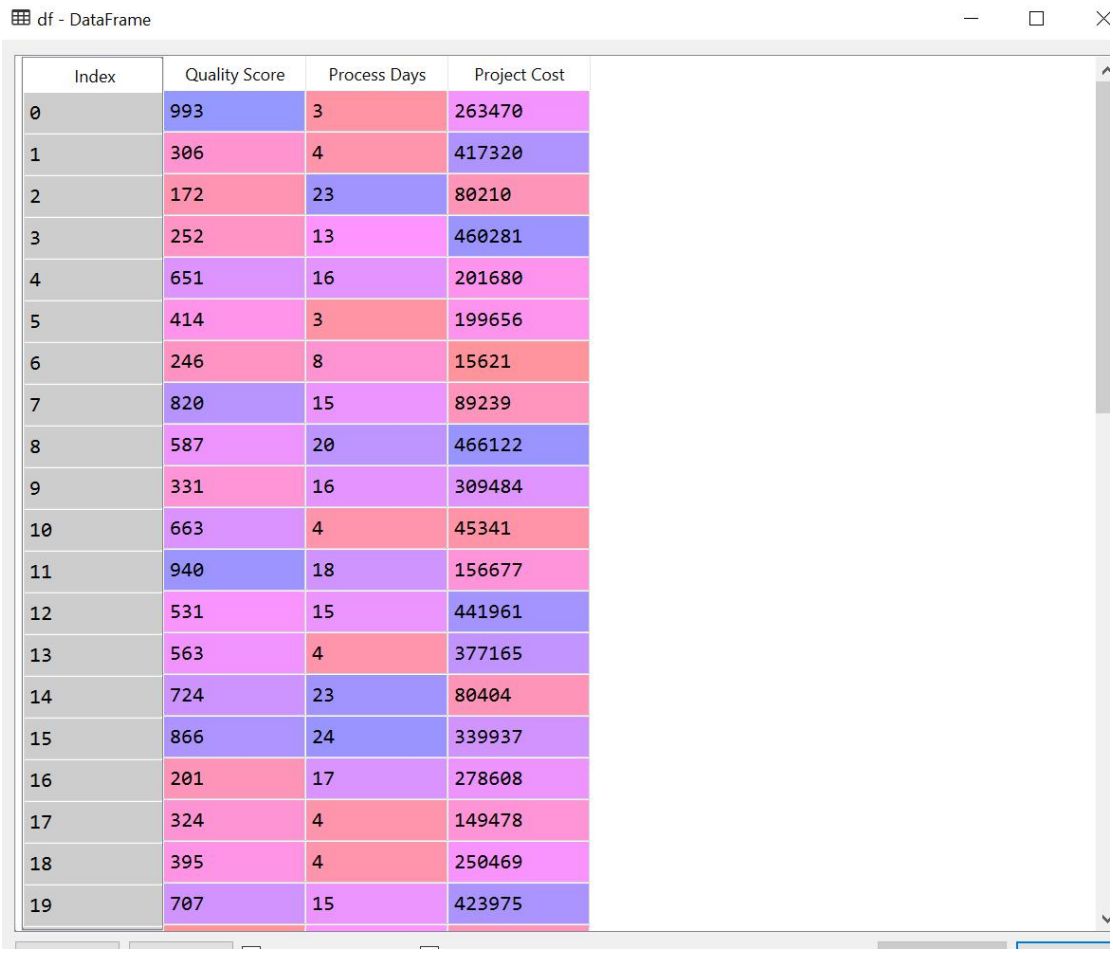
**Analysis**

I choose Python over R to do this project as Python is easy to write clean and tidy function. Python is an alternative choice for R. As for the problem, it is easy to first do some small calculation first before answering the question. I used Python to calculate the number of score of each project and adopted Excel to post process the data.

First, I import "pandas" and give a name called "pd". I import "Excelwriter" from "pandas" and "numpy" library. I used a function called "read_excel" from pd to read the raw data and assign to the variable "raw_data". I created a data frame called"df" and there are only three columns which are "Quality Score", "Process Days", "Project Cost". They are assigned to "xs", "ys", "zs", respectively.

```
import pandas as pd
from pandas import ExcelWriter
import numpy as np
raw_data=pd.read_excel('Module 2 Project_Project Performance_v1(2).xlsx')
df= pd.DataFrame(raw_data, columns=['Quality Score','Process Days','Project Cost'])
#print df
xs= df['Quality Score']
ys= df['Process Days']
zs= df['Project Cost']
```

The data frame "df" is shown below.




| Index | Quality Score | Process Days | Project Cost |
|---|---|---|---|
| 0 | 993 | 3 | 263470 |
| 1 | 306 | 4 | 417320 |
| 2 | 172 | 23 | 80210 |
| 3 | 252 | 13 | 460281 |
| 4 | 651 | 16 | 201680 |
| 5 | 414 | 3 | 199656 |
| 6 | 246 | 8 | 15621 |
| 7 | 820 | 15 | 89239 |
| 8 | 587 | 20 | 466122 |
| 9 | 331 | 16 | 309484 |
| 10 | 663 | 4 | 45341 |
| 11 | 940 | 18 | 156677 |
| 12 | 531 | 15 | 441961 |
| 13 | 563 | 4 | 377165 |
| 14 | 724 | 23 | 80404 |
| 15 | 866 | 24 | 339937 |
| 16 | 201 | 17 | 278608 |
| 17 | 324 | 4 | 149478 |
| 18 | 395 | 4 | 250469 |
| 19 | 707 | 15 | 423975 |

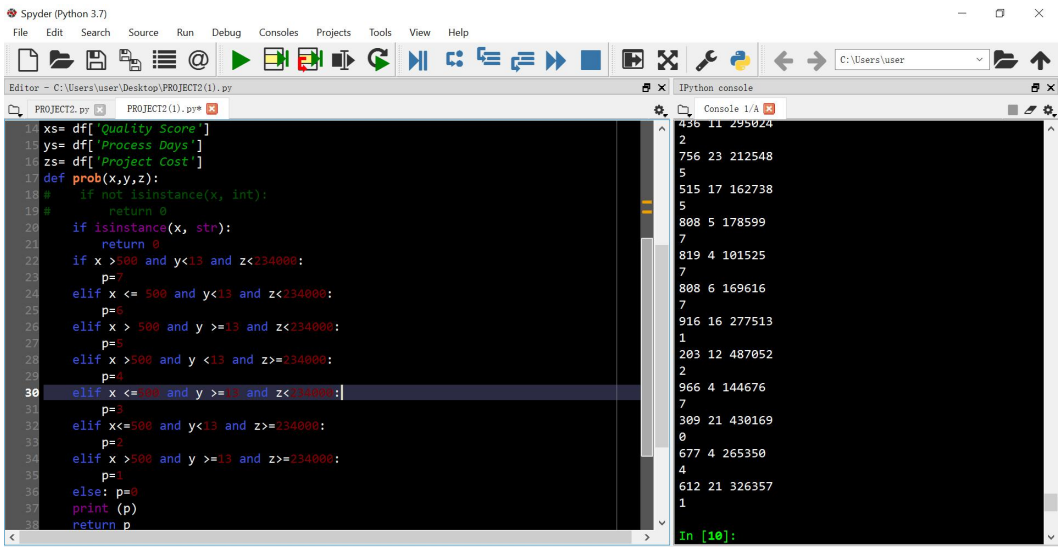
I want to get a score called “p” for each row. I created a function using key word “def” and call it “prob”. There are three input values-----x,y,z. Then I used if, elif and else statement to examine if the statement is satisfied, then a p score is assigned. If not satisfied, then go to another statement. At the end of the function, print and return value p.

```
def prob(x,y,z):
#    if not isinstance(x, int):
#        return 0
    if isinstance(x, str):
        return 0
    if x >500 and y<13 and z<234000:
        p=7
    elif x <= 500 and y<13 and z<234000:
        p=6
    elif x > 500 and y >=13 and z<234000:
        p=5
    elif x >500 and y <13 and z>=234000:
        p=4
    elif x <=500 and y >=13 and z<234000:
        p=3
    elif x<=500 and y<13 and z>=234000:
        p=2
    elif x >500 and y >=13 and z>=234000:
        p=1
    else: p=0
    print (p)
    return p
```

The next function is to go through all the rows and append a “p” score at the end. I utilized “for loop” to realize it. In here, “row” denotes an index within the range of length “xs”, which is 50. Then print values in each row and conduct the function created before called “prob’. The return value p is added in column called “p”.

```
for row in range(len(xs)):
    print (xs[row],ys[row],zs[row])
#    df['p']= prob(row['Quality Score'],row['Process Days'],row['Project Cost'])
    df['p']= prob(xs[row],ys[row],zs[row])
```

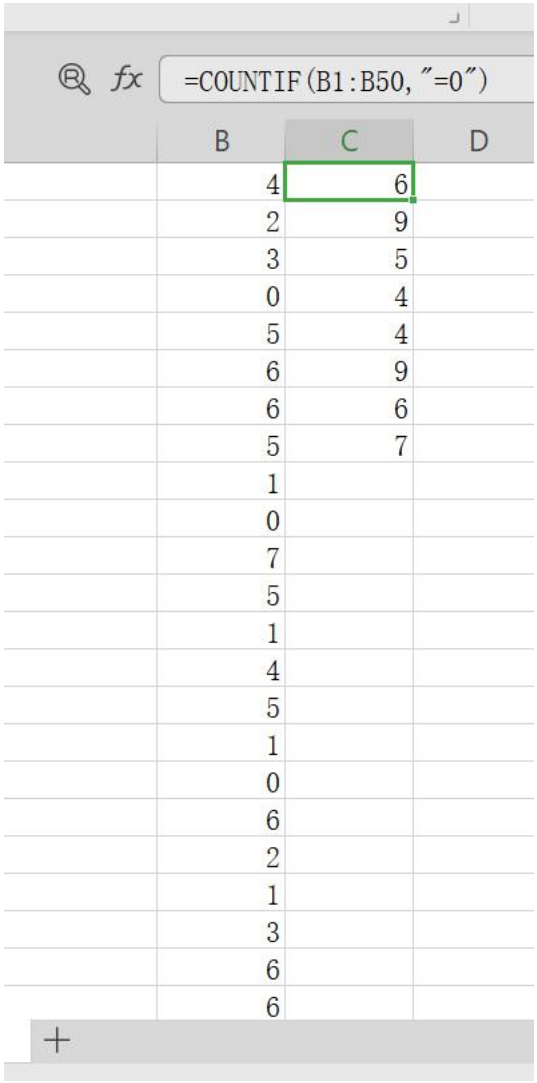
As shown below, on the right the console displays each score, which is a number range from 0 to 7.

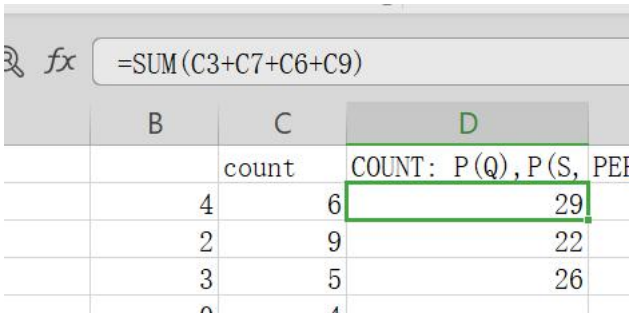


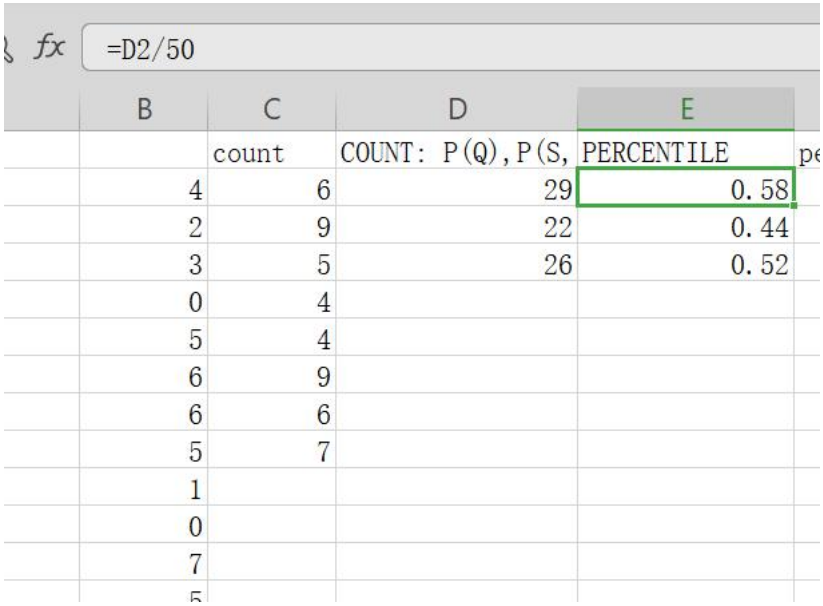

On the second part, I export the score of each project and input in Excel spreadsheet. In Excel the function is easy to calculate the probability.

1. Calculalte the value, count and probability of P(Q),P(S), P(C).

I used “COUONTIF” function to calculate for each score, what is the frequency each score occurs. And sum the total count for score P,Q,C. The count total for each P,Q,C divided by 50 is the probability.

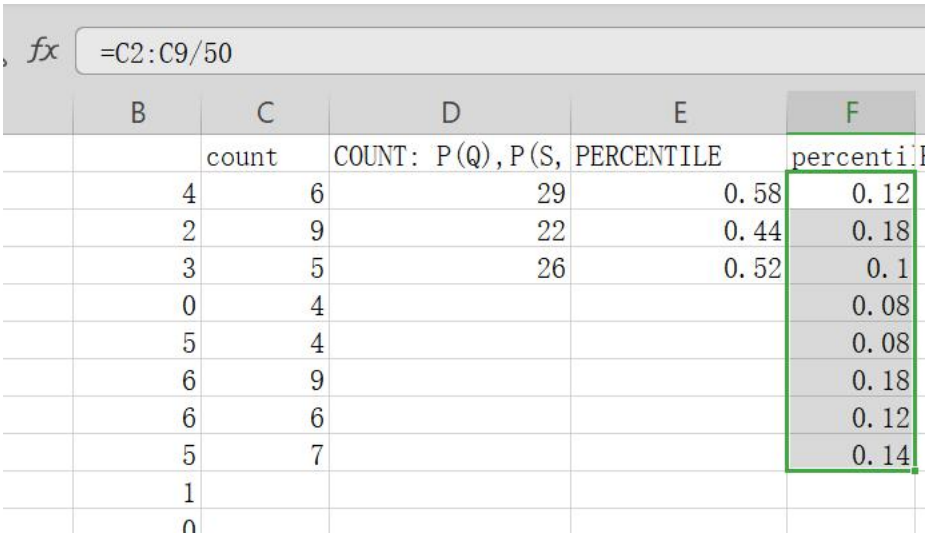




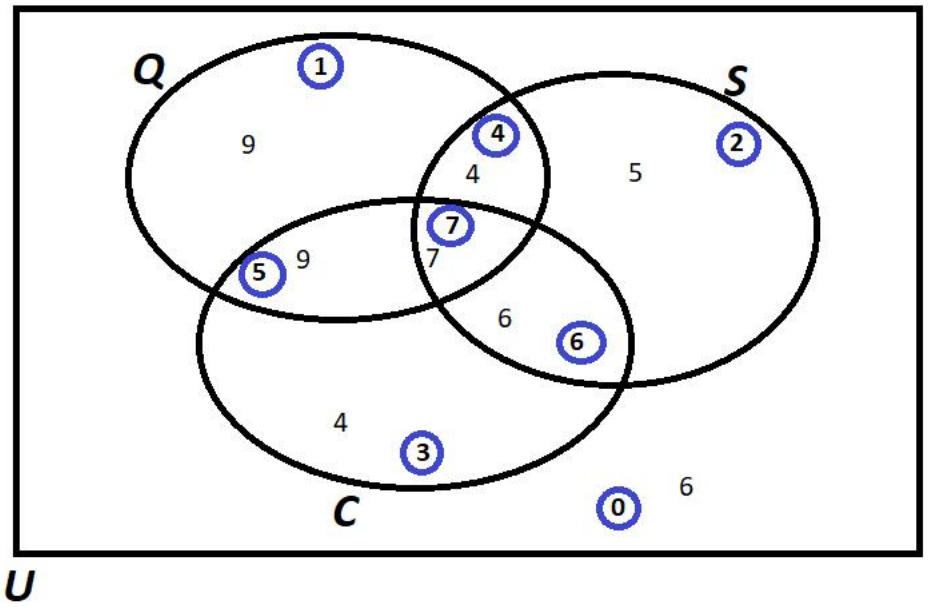

| Notation for the Probability of the given event: | Count: | Probability/Percentile |
|---|---|---|
| P(Q) | 29 | 0.58 |
| P(S) | 22 | 0.44 |
| P(C) | 26 | 0.52 |

2. Calculate total number counted for each score and the probability.
I divide each count in Colume “C” by 50.



| | Notation for the Probability of the given event in terms of events Q, S, and C: | Count: | Probability/ Percentile |
|---|---|---|---|
| P(Score=0) | $P(Q')\cap P(S')\cap P(C')$ | 6 | 0.12 |
| P(Score=1) | $P(Q)\cap(P(S')\cap P(C'))$ | 9 | 0.18 |
| P(Score=2) | $P(S)\cap(P(Q')\cap P(C'))$ | 5 | 0.1 |
| P($S_c$ore=3) | $P(C)\cap(P(Q')\cap P(S'))$ | 4 | 0.08 |
| P(Score=4) | $(P(Q)\cap(P(S))\cap P(C')$ | 4 | 0.08 |
| P(Score=5) | $(P(Q)\cap(P(C))\cap P(S')$ | 9 | 0.18 |
| P(Score=6) | $(P(S)\cap(P(C))\cap P(Q')$ | 6 | 0.12 |
| P(Score=7) | $(P(S)\cap(P(C))\cap P(Q)$ | 7 | 0.14 |

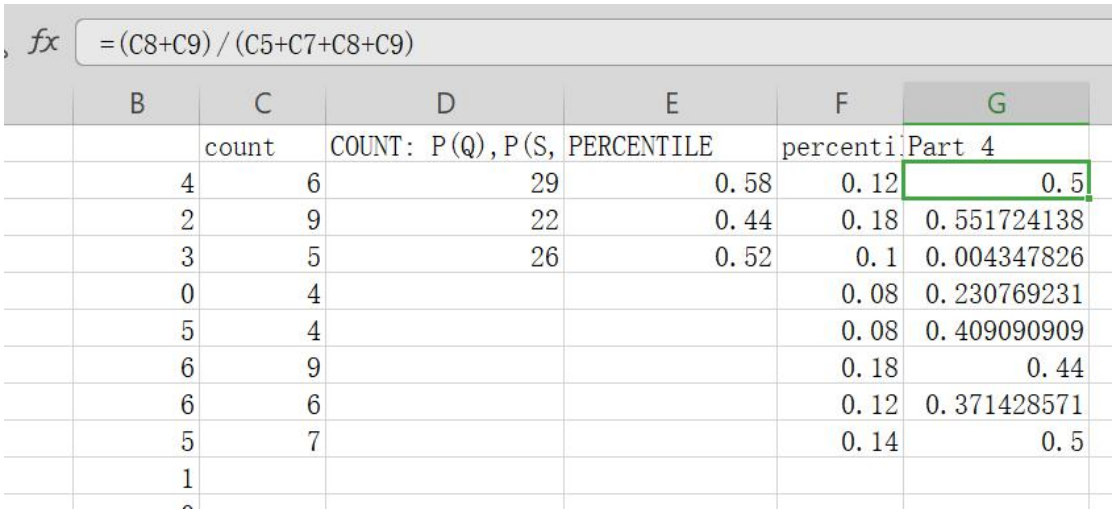
3. The Venn diagram of this project management procedure consists of 8 regions that have been numbered 0 through 7 and according to the score gained by the project. Use the Venn diagram provided in the Excel workbook by first copying the Venn diagram into Microsoft Paint and then using Paint to populate the eight regions with the counts of the projects that are located in each region. Copy the resulting Venn diagram into your report.




4.

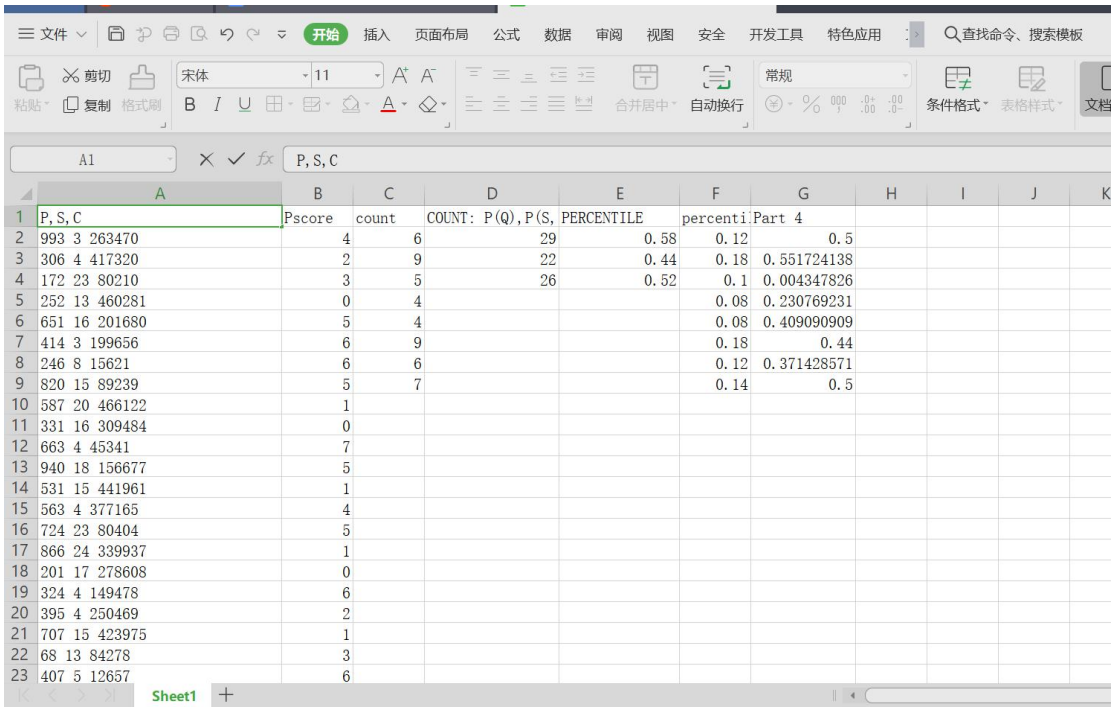
In the word document, write the mathematical notation and the formula for calculating the probabilities of the events outlined below. Notation for the Probability of the given event in terms of events Q, S, and C: Count: Probability/Percentage P(Score = 0) P(Score = 1) P(Score = 2) P(Score = 3) P(Score = 4) P(Score = 5)

P(Score = 6) P(Score = 7)
Write the mathematical notation for the event in your report;
Using R, calculate the total counts for each event;
Using R, calculate the probability (or percentage) for each event.
The events are: a) Of those who satisfied Cost, what percentage also satisfied Speed? b) Of those who satisfied Quality, what percentage also satisfied Cost? c) Of those who satisfied Quality, what percentage also satisfied Speed but did not satisfy the Cost? d) Of those who satisfied Cost, what percentage also satisfied Speed but did not satisfy the Quality? e) Of those who did not satisfy Speed, what percentage satisfied Quality and Cost? f) What percentage satisfied exactly two of the three criteria? g) Of those who satisfied at least one of the three criteria, what percentage satisfied exactly one criterion? h) Of those who did not satisfy Cost, what percentage satisfied the Speed criterion?




| | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| | | count | COUNT: P(Q), P(S, | PERCENTILE | percenti | Part 4 |
| | 4 | 6 | 29 | 0.58 | 0.12 | 0.5 |
| | 2 | 9 | 22 | 0.44 | 0.18 | 0.551724138 |
| | 3 | 5 | 26 | 0.52 | 0.1 | 0.004347826 |
| | 0 | 4 | | | 0.08 | 0.230769231 |
| | 5 | 4 | | | 0.08 | 0.409090909 |
| | 6 | 9 | | | 0.18 | 0.44 |
| | 6 | 6 | | | 0.12 | 0.371428571 |
| | 5 | 7 | | | 0.14 | 0.5 |
| | 1 | | | | | |

| | Notation for the Probability of the given event | Count: | Probability/ Percentile |
|---|---|---|---|
| (a) | $P(S=6)\cup P(S=7) \mid P(S=3)\cup P(S=5)\cup P(S=6)\cup P(S=7)$ | 13 | 0.5 |
| (b) | $P(S=5)\cup P(S=7) \mid P(S=1)\cup P(S=4)\cup P(S=5)\cup P(S=7)$ | 16 | 0.551724138 |
| (c) | $P(S=4) \mid P(S=1)\cup P(S=4)\cup P(S=5)\cup P(S=7)$ | 4 | $0.\dot{0}04347826$ |
| (d) | $P(S=6) \mid P(S=3)\cup P(S=5)\cup P(S=6)\cup P(S=7)$ | 6 | 0.230769231 |
| (e) | $P(S=5) \mid P(S=1)\cup P(S=3)\cup P(S=5)$ | 9 | 0.409090909 |
| (f) | $P(S=5)\cup P(S=6)\cup P(S=7)/50$ | 22 | 0.44 |
| (g) | $P(S=2)\cup P(S=3)\cup P(S=4) \mid P(S=2)\cup P(S=PP3)\cup P(S=4)\cup P(S=5)\cup P(S=6)\cup P(S=7)$ | 13 | 0.371428571 |
| (h) | $P(S=2)\cup P(S=4) \mid P(S=1)\cup P(S=2)\cup P(S=4)$ | 9 | $0.\dot{5}$ |



| P, S, C | Pscore | count | COUNT: P(Q), P(S, | PERCENTILE | percenti | Part 4 |
|---|---|---|---|---|---|---|
| 993 3 263470 | 4 | 6 | 29 | 0.58 | 0.12 | 0.5 |
| 306 4 417320 | 2 | 9 | 22 | 0.44 | 0.18 | 0.551724138 |
| 172 23 80210 | 3 | 5 | 26 | 0.52 | 0.1 | 0.004347826 |
| 252 13 460281 | 0 | 4 | | | 0.08 | 0.230769231 |
| 651 16 201680 | 5 | 4 | | | 0.08 | 0.409090909 |
| 414 3 199656 | 6 | 9 | | | 0.18 | 0.44 |
| 246 8 15621 | 6 | 6 | | | 0.12 | 0.371428571 |
| 820 15 89239 | 5 | 7 | | | 0.14 | 0.5 |
| 587 20 466122 | 1 | | | | | |
| 331 16 309484 | 0 | | | | | |
| 663 4 45341 | 7 | | | | | |
| 940 18 156677 | 5 | | | | | |
| 531 15 441961 | 1 | | | | | |
| 563 4 377165 | 4 | | | | | |
| 724 23 80404 | 5 | | | | | |
| 866 24 339937 | 1 | | | | | |
| 201 17 278608 | 0 | | | | | |
| 324 4 149478 | 6 | | | | | |
| 395 4 250469 | 2 | | | | | |
| 707 15 423975 | 1 | | | | | |
| 68 13 84278 | 3 | | | | | |
| 407 5 12657 | 6 | | | | | |

## Conclusion

1) The largest frequency of score is two and five, then followed by seven, meaning these three number of score has large probability if randomly choosing a project to examine.
2) As see in Venn Picture, some event occurs by correlation. For example, for a project, the event could be Quality is satisfied while other two are not satisfied. We call it conditional probability.
3) We can examine in many layers, for example the percentile of Quality is satisfied which is equal to 0.58%. Compared to Speed and Cost, Quality that pass the standard has larger probability than Speed or Cost.

## Reference

[1] Bluman, A. G. (2009). Elementary statistics: A step by step approach. New York: McGraw-Hill Higher Education