Final Report

Group4

1.Motivation and Questions

Data is everywhere today, and data Science is a field that manages, researches, analyzes data, and thus provides meaningful information to non-professionals. It's one of the hottest professions of the 21st century. As a result, data scientists are in high demand and receive competitive salaries. However, the salaries for data scientists may vary widely due to a range of factors. Understanding these variations is crucial for aspiring data scientists. For students planning their career path, knowing which factors most influence salary can help them make informed decisions about their skill development and education. In this project, we aim to analyze the factors, such as the level of formal education, company type, and the number of proficient languages, that affect the salary of a data scientist. By identifying these patterns, we hope to provide valuable insights and practical reference for those in need. Here are the issues we aim to explore :

a. Does studying a master degree have a great impact on salaries?
b. Does proficiency in many languages (3 or more) result in higher salaries?
c. Does the work location have a significant impact on employees' average salaries?

2.Data Collection

a. We downloaded an existing dataset from Kaggle, which was collected using a questionnaire. The dataset consists of 34 questions, and each column is labeled with the type of response expected for that particular question.

b. In order to analyze the above problems, we organize the ‘Education_Level’, 'Programming_Languages', and 'Country' columns as follow:

Education_Level---

Higher Education: ['Doctoral degree', 'Master']

Lower Education: ['Bachelor', 'Professional degree', 'High School']

—---

Programing_Languages—---

Type1: proficient in three languages or more.

Type2: proficient in two languages or less.

—---

Country—---

Europe: ['France','Germany','Netherlands','Ireland','Greece','Ukraine','Belarus', 'United Kingdom of Great Britain and Northern Ireland' ,'Sweden' ,'Portugal', 'Poland','Italy','Czech Republic', 'Spain','Hungary','Norway','Switzerland', 'Denmark', 'Romania' ,'Belgium','Austria']

Asia: ['India','Australia','Russia','Pakistan','Japan','South Korea', 'Indonesia', 'Hong Kong (S.A.R.)','Turkey','Singapore','Israel','Taiwan', 'Bangladesh', 'Thailand','China','Viet Nam','Republic of Korea','New Zealand','Malaysia' ,'Philippines','Saudi Arabia','Iran, Islamic Republic of...']

America: ['United States of America','Brazil','Mexico','Canada', 'Chile', 'Argentina' ,'Colombia','Peru']

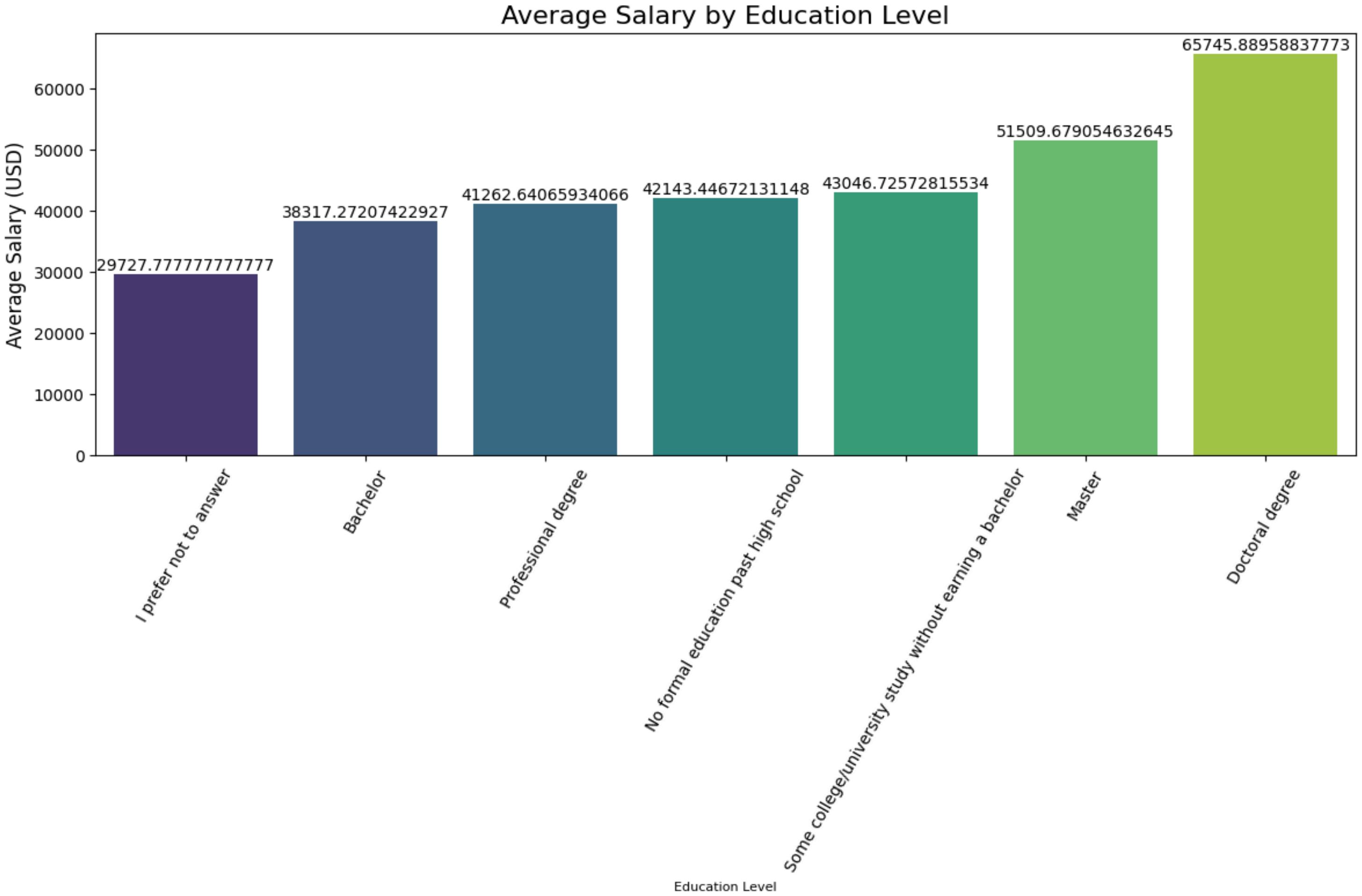
Africa: ['Nigeria','Morocco','South Africa','Egypt','Tunisia','Kenya','Algeria']

—---

c. After data preprocessing, we still have 15,703 data entries available for analysis.
d. After data preprocessing, there still exists gender imbalance among survey participants, with 12,952 males and 2,448 females. Respondents with lower salaries might be less willing to disclose their salary information compared to those with higher salaries. These factors may lead to discrepancies between the analysis results and the actual facts.

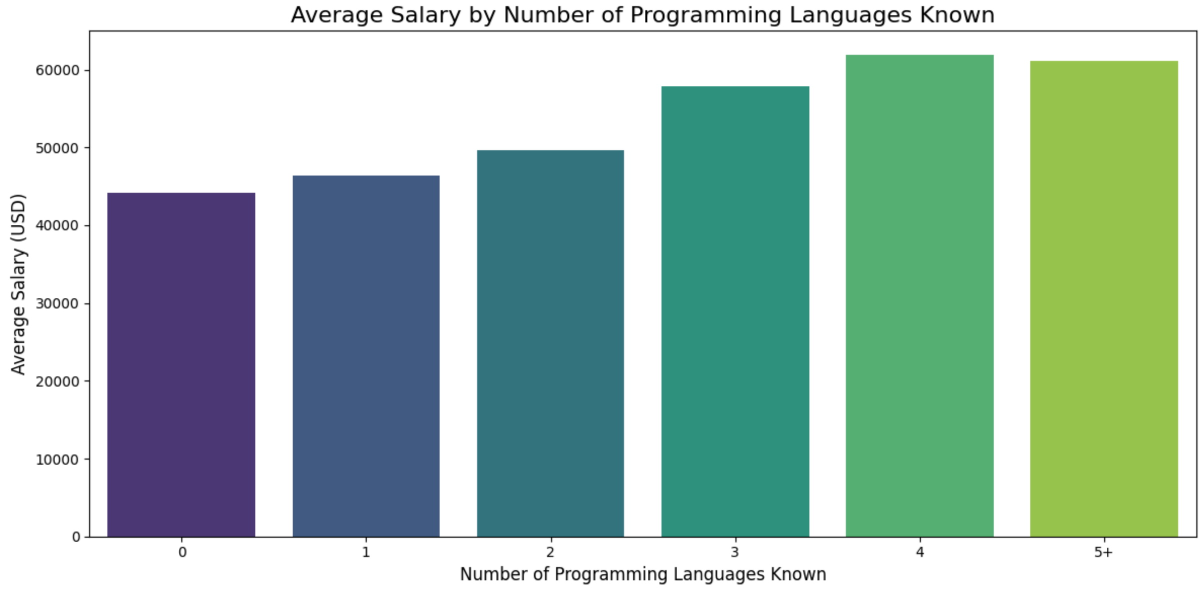
3.Descriptive Analysis

Does studying a master degree have a great impact on salaries?



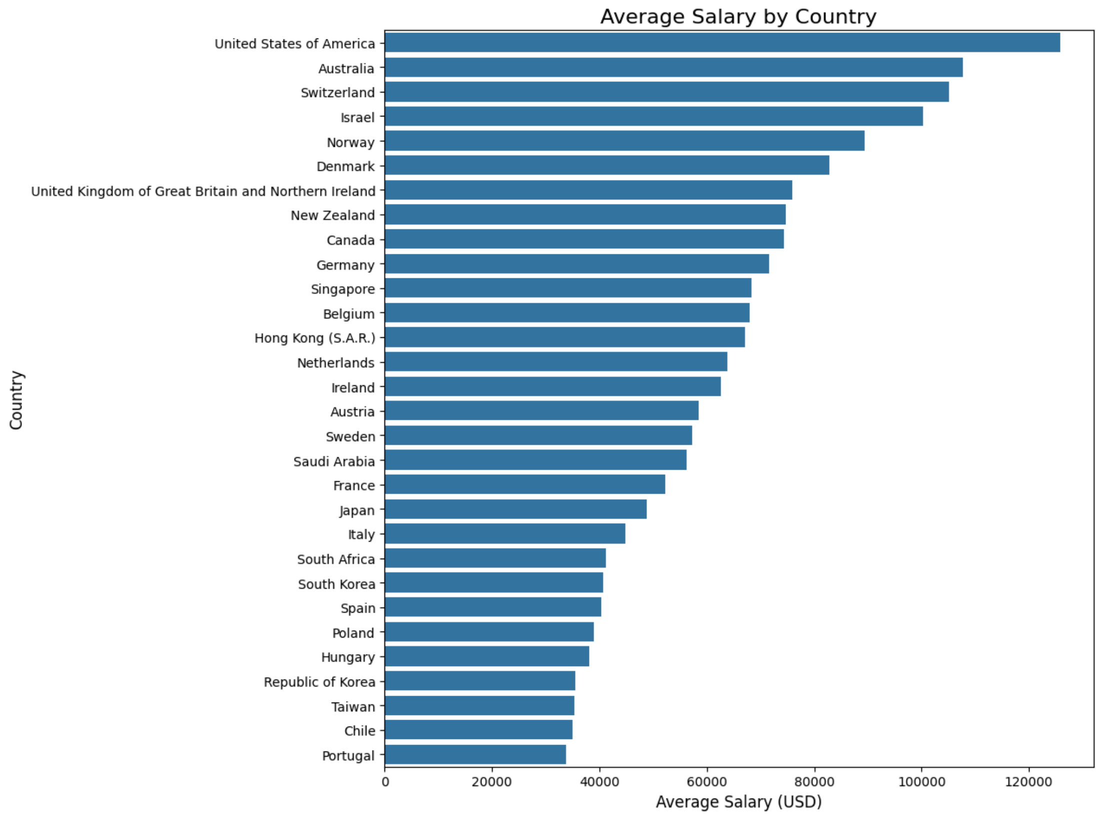
The bar chart “Average Salary by Education Level” indicates that individuals with a master’s degree earn a significantly high average salary compared to those with only a bachelor's degree or less education.

How many programming languages do software engineers have to code(know) if they want to earn a decent salary?



Based on the bar chart “Average Salary by Number of Programming Languages Known”, we can state that the more programming languages you know, there is a high chance that you will earn more money. We find that there are a lot of low salaries occurring in people who know around 7-9 languages. However, there’s only a few entries in high numbers of languages, so we merge them into category 5+.

Does the work location have a significant impact on employees’ average salaries?



Obviously, we can observe that the employees in countries like the United States, Switzerland earn substantially higher average salaries, indicating that work location has a major impact on salary levels. The table above only shows the first 30 countries in the order.

4.Statistical Test

a. Does studying a master degree have a great impact on salaries? (high_edu sample : 7951 / low_edu sample : 3823)

   $\mu 1$ = The mean salary for individuals with a master's degree or higher.

   $\mu 2$ = The mean salary for individuals with a bachelor's degree or lower.

   H0: $\mu 1 = \mu 2$

   Ha: $\mu 1 > \mu 2$

```
high education level salary mean:  57858.166016853225
high education level salary variance:  4926568586.818986
low education level salary mean:  41928.46128694742
low education level salary variance:  4432385283.804841
```

   Step 1 : use F-test to determine which t-test to apply (with 5 % sig. level)

   $\sigma 1$ = The mean salary for individuals with a bachelor's degree or lower.

   $\sigma 2$ = The mean salary for individuals with a master's degree or higher.

   H0: $\sigma 1^2 = \sigma 2^2$

   Ha: $\sigma 1^2 \neq \sigma 2^2$

```
F: 1.1114937604408632
accept range: 0.9471478375314608 ~ 1.0563602312906089
False
```

   Step 2 : using Welch's t-test (with 5 % sig. level)

```
degree of freedom: 7912
t0:  11.943120136680323
t_value:  1.645046239269868
True
```

   => since t0 is larger than the critical value, we have evidence to reject H0 at alpha = 0.05 and conclude that $\mu 1 > \mu 2$.

   Conclusion: The mean salary of individuals with a master's degree or higher is not equal to the mean salary of individuals with a bachelor's degree or lower

b. Does proficiency in more than two languages result in a salary increase? (sample size : 12497)

$\mu 1$ = The mean salary for individuals proficient in three languages or more.

$\mu 2$ = The mean salary for individuals proficient in two languages or less.

H0: $\mu 1 = \mu 2$

Ha: $\mu 1 > \mu 2$

```
type1 mean:  59564.90380094044
type2 mean:  47211.046665764916
type1 variance:  5215053585.598015
type2 variance:  4564480840.815622
```

Step 1 : use F-test to determine which t-test to apply (with 5 % sig. level)

σ1 = The standard devation of salary for individuals proficient in three languages or more.

σ2 = The standard devation of salary for individuals proficient in two languages or less.

H0: $\sigma 1^2 = \sigma 2^2$

Ha: $\sigma 1^2 \neq \sigma 2^2$

```
F : 1.1425294064036737
accept range: 0.9506870925274047 ~ 1.0516222256853824
False
```

Step 2 : using Welch's t-test (with 5 % sig. level)

```
degree of freedom:  8012
t0:  9.083100489001893
t_value:  1.6450438349422825
True
```

=> Since t0 is larger than the critical value, we have evidence to reject H0 at alpha = 0.05 and conclude that $\mu 1 > \mu 2$

Conclusion : The mean salary for individuals proficient in three languages or more is higher than the mean salary for individuals proficient in two languages or less.

c. Does the mean salary vary in different continents?

$\mu 1$ = The mean salary for individuals in Europe.

$\mu 2$ = The mean salary for individuals in Asia.

$\mu 3$ = The mean salary for individuals in America.

$\mu 4$ = The mean salary for individuals in Africa.

H0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$

Ha: At least one mean salary of a continent is different.

```
mean :  EU:54730.28720525705 AS:31201.7107398107 AME:91917.66180294713 AF:15776.455026455027
var :   EU:2625423121.116806 AS:3192872507.100496 AME:7278473547.945596 AF:1707493335.997532
F value :622.9168378377203 critical_value :2.605662391789357
True
```

=> since F is larger than the critical value, we have evidence to reject H0 at alpha = 0.05.

Conclusion : At least one mean salary of a continent is different.

Bonferroni Method

$\mu 1$ = The mean salary for individuals in Europe.

$\mu 2$ = The mean salary for individuals in Asia.

$\mu 3$ = The mean salary for individuals in America.

$\mu 4$ = The mean salary for individuals in Africa.

H0: $\mu 1 = \mu 2$

H1: $\mu 1 = \mu 3$

H2: $\mu 1 = \mu 4$

H3: $\mu 2 = \mu 3$

H4: $\mu 2 = \mu 4$

H5: $\mu 3 = \mu 4$

```
H0 interval: 19453.43181964929 ~ 27603.721111243416
H1 interval: -41586.27814194457 ~ -32788.47105343558
H2 interval: 31105.47675127508 ~ 46802.18760632897
H3 interval: -64432.19556504162 ~ -56999.70656123123
H4 interval: 7938.146543383758 ~ 22912.364883327587
H5 interval: 68473.06800423688 ~ 83809.34554874731
```

Conclusion：$\mu_1 \neq \mu_2$, $\mu_1 \neq \mu_3$, $\mu_1 \neq \mu_4$, $\mu_2 \neq \mu_3$, $\mu_2 \neq \mu_4$, $\mu_3 \neq \mu_4$.

5.Response to Peer Review Comments

a. 地區則可以在分析結果時與該地區薪資中位數來比較作出合理的結論。
   Ans : We are going to discuss the difference between countries for data scientists. Instead of the difference between data scientists and other jobs in a specific region.
b. 假設中沒有詳細提到要取欄位中哪些類別的資料比較(像是最高學歷欄位取"碩士"與"高中"進行比較，或是more programming languages指的語言數量為多少)
   Ans : In final project, we further differentiated the columns to be compared (eg. education level : we categorized respondents into those of master’s degree or higher and those without; for programming languages: we used two programming languages as classification benchmark.)
c. 樣本數量>30，應該三個假設檢定都要使用z-test
   Ans : With sufficiently large sample size (15,703 here), the t-test approaches the z-test, so we can also use the t-test for the analysis.
d. The effect of the use of LLM like Chat GPT, Claude, Gemini, etc. might affect the salary of a Data Scientist.
   Ans : This survey was collected in 2019, when LLMs were not yet popular. Therefore, this analysis is based on a time before LLMs had been fully developed. If LLMs were to be considered, the questionnaire may need to be redesigned, and a new round of responses would be required.
e. It’s unclear whether the data contains salary information for the same individuals or groups from the same industry and region before and after COVID-19. If such comparisons are available, a paired t-test would be appropriate; otherwise, an independent t-test would be more suitable.

Ans : After revisiting the dataset, we realized that it is not feasible to compare pre- and post-COVID-19 scenarios based on the existing data in the questionnaire. Therefore, we abandoned this question and shifted our focus to analyzing the relationship between salary and other variables.

6.Creativity and Others

a. Creativity

First, since the dataset contains several missing values and duplicate rows, we carefully clean the data by removing any inappropriate entries. Second, to effectively test our hypotheses, we employ various grouping strategies, such as grouping countries by their continents. This approach successfully not only helps us to simplify our analysis but enables us to make accurate hypotheses.

b. Code for three hypotheses:

```
"""### Q1: Does having an education level higher than a bachelor’s
degree result in a salary increase?
μ1 = The mean salary for individuals with a master’s degree or higher.
<br>
μ2 = The mean salary for individuals with a bachelor’s degree or lower.
##### H0: μ1 = μ2
##### Ha: μ1 > μ2
"""

education_categories = {
    'Higher Education': ['Doctoral degree', 'Master'],
    'Lower Education': ['Bachelor', 'Professional degree', 'High
School']
}

high_edu = df[df['Education Level'].isin(education_categories['Higher
Education'])]
high_mean = high_edu['Salary_numeric'].mean()
high_variance = high_edu['Salary_numeric'].var()

# Filter Lower Education
```

```
low_edu = df[df['Education Level'].isin(education_categories['Lower
Education'])]
low_mean = low_edu['Salary_numeric'].mean()
low_variance = low_edu['Salary_numeric'].var()

print("high education level salary mean: ", high_mean)
print("high education level salary variance: ", high_variance)
print("low education level salary mean: ", low_mean)
print("low education level salary variance: ", low_variance)

"""#### Step1: First use F-test to test which t-test to apply.(at 5%
significance level)
σ1 = The mean salary for individuals with a bachelor’s degree or lower.
<br>
σ2 = The mean salary for individuals with a master’s degree or higher.
##### H0: $σ1^2$ = $σ2^2$
##### Ha: $σ1^2$ ≠ $σ2^2$
"""

print(len(high_edu))
print(len(low_edu))

import scipy.stats as stats
alpha = 0.05
F = high_variance / low_variance
v1, v2 = len(high_edu) - 1, len(low_edu) - 1

f_upper = stats.f.ppf(1 - alpha / 2, v1, v2)
f_lower = 1 / stats.f.ppf(1 - alpha / 2, v2, v1)

print(f'F: {F}')
print(f'accept range: {f_lower} ~ {f_upper}')
print(f_upper >= F and F >= f_lower)

"""### Step2: Since we do not have enough evidence to support the
variance of two distribution is equal, we should use Welch's
t-test.(one sided, at 5% significance level)


"""

import math
n1 = len(high_edu)
```

```
n2 = len(low_edu)
var1 = high_variance
var2 = low_variance
mean1 = high_mean
mean2 = low_mean
delta = 0
alpha = 0.05
degree_of_freedom = math.floor(((var1 / n1) + (var2 / n2)) ** 2 /
((var1 / n1) ** 2 / (n1 - 1) + (var2 / n2) ** 2 / (n2 - 1)))
t0 = (mean1 - mean2 - delta) / math.sqrt((var1 / n1) + (var2 / n2))

t_value = stats.t.ppf(1 - alpha, degree_of_freedom)
# p_value = stats.t.sf(abs(t0), degree_of_freedom) * 2
# print(degree_of_freedom, t0, t_value)
# print(t0 > t_value)
print(f"degree of freedom: {degree_of_freedom}")
print("t0: ", t0)
print("t_value: ", t_value)
print(t0 > t_value)

"""### Since t0 is larger than the critical value, we have evidence to
reject H0 at alpha = 0.05 and conclude that μ1 > μ2.

### Q2: Does proficient in more than two language result in a salary
increase?
μ1 = The mean salary for individuals proficient in three languages or
more. <br>
μ2 = The mean salary for individuals proficient in two languages or
less.
##### H0: μ1 = μ2
##### Ha: μ1 > μ2
"""

df['Programming_Languages'] = df['Programming_Languages'].apply(
    lambda x: x if isinstance(x, list) else []
)

lans = df['Programming_Languages']
salaries = df['Salary_numeric'].tolist()

print(lans.shape)
```

```
type1, type2 = [], []
sal1, sal2 = [], []
for i, lan in enumerate(lans):
  if len(lan) >= 3:
    type1.append(lan)
    sal1.append(salaries[i])

  else:
    type2.append(lan)
    sal2.append(salaries[i])
sal1, sal2 = np.array(sal1), np.array(sal2)

type1_mean = sal1.mean()
type2_mean = sal2.mean()
type1_var = sal1.var()
type2_var = sal2.var()

#type1: three or more languages
#type2: 2 or less languages
print("type1 mean: ", type1_mean)
print("type2 mean: ", type2_mean)
print("type1 variance: ", type1_var)
print("type2 variance: ", type2_var)

"""#### Step1: First use F-test to test which t-test to apply.(at 5%
significance level)
σ1 = The standard devation of salary for individuals proficient in
three languages or more.<br>
σ2 = The standard devation of salary for individuals proficient in two
languages or less.
##### H0: $σ1^2$ = $σ2^2$
##### Ha: $σ1^2$ ≠ $σ2^2$
"""

import scipy.stats as stats
alpha = 0.05
F = type1_var / type2_var
v1, v2 = len(type1) - 1, len(type2) - 1

f_upper = stats.f.ppf(1 - alpha / 2, v1, v2)
f_lower = 1 / stats.f.ppf(1 - alpha / 2, v2, v1)

print(f'F : {F}')
```

```python
print(f'accept range: {f_lower} ~ {f_upper}')
print(f_upper >= F and F >= f_lower)

"""### Step2: Since we do not have enough evidence to support the
variance of two distribution is equal, we should use Welch's
t-test.(one sided, at 5% significance level)"""

import math
n1 = len(high_edu)
n2 = len(low_edu)
var1 = type1_var
var2 = type2_var
mean1 = type1_mean
mean2 = type2_mean

delta = 0
alpha = 0.05
degree_of_freedom = math.floor(((var1 / n1) + (var2 / n2)) ** 2 /
((var1 / n1) ** 2 / (n1 - 1) + (var2 / n2) ** 2 / (n2 - 1)))
t0 = (mean1 - mean2 - delta) / math.sqrt((var1 / n1) + (var2 / n2))

t_value = stats.t.ppf(1 - alpha, degree_of_freedom)
# p_value = stats.t.sf(abs(t0), degree_of_freedom) * 2
# print(degree_of_freedom, t0, t_value)
# print(t0 > t_value)
print("degree of freedom: ", degree_of_freedom)
print("t0: ", t0)
print("t_value: ", t_value)
print(t0 > t_value)

"""### Since t0 is larger than the critical value, we have evidence to
reject H0 at alpha = 0.05 and conclude that μ1 > μ2.

### Q3: Does the mean salary vary in different continents ?
μ1 = The mean salary for individuals in Europe. <br>
μ2 = The mean salary for individuals in Asia. <br>
μ3 = The mean salary for individuals in America. <br>
μ4 = The mean salary for individuals in Africa.
##### H0: μ1 = μ2 = μ3 = μ4
##### Ha: At least one mean salary of a continent is different.
"""

import scipy.stats as stats
```

```
Europe=['France',
'Germany','Netherlands','Ireland','Greece','Ukraine','Belarus','United
Kingdom of Great Britain and Northern Ireland','Sweden','Portugal'
,'Poland','Italy','Czech
Republic','Spain','Hungary','Norway','Switzerland','Denmark','Romania',
'Belgium','Austria']
Asia=['India','Australia','Russia','Pakistan','Japan','South
Korea','Indonesia','Hong Kong
(S.A.R.)','Turkey','Singapore','Israel','Taiwan','Bangladesh','Thailand
','China'
,'Viet Nam','Republic of Korea','New
Zealand','Malaysia','Philippines','Saudi Arabia','Iran, Islamic
Republic of...']
America=['United States of
America','Brazil','Mexico','Canada','Chile','Argentina','Colombia','Per
u']
Africa=['Nigeria','Morocco','South
Africa','Egypt','Tunisia','Kenya','Algeria']
Others = ['Other']

all = Europe + Asia + America + Africa + Others

df_europe = df[df['Country'].isin(Europe)]
df_asia = df[df['Country'].isin(Asia)]
df_america = df[df['Country'].isin(America)]
df_africa = df[df['Country'].isin(Africa)]

mean_df_europe = df_europe['Salary_numeric'].mean()
mean_df_asia = df_asia['Salary_numeric'].mean()
mean_df_america = df_america['Salary_numeric'].mean()
mean_df_africa = df_africa['Salary_numeric'].mean()
var_df_europe = df_europe['Salary_numeric'].var()
var_df_asia = df_asia['Salary_numeric'].var()
var_df_america = df_america['Salary_numeric'].var()
var_df_africa = df_africa['Salary_numeric'].var()

print("mean: ", mean_df_europe, mean_df_asia, mean_df_america,
mean_df_africa)
print("var: ", var_df_europe, var_df_asia, var_df_america,
var_df_africa)

mean_all = sum(df[df['Country'] != 'Other']['Salary_numeric']) /
len(df[df['Country'] != 'Other'])
```

```
SSE = (df_europe.shape[0] - 1) * var_df_europe + (df_asia.shape[0] - 1)
* var_df_asia + (df_america.shape[0] - 1) * var_df_america
SST = (mean_df_europe - mean_all) ** 2 * df_europe.shape[0] +
(mean_df_asia - mean_all) ** 2 * df_asia.shape[0] + (mean_df_america -
mean_all) ** 2 * df_america.shape[0]
dft = 3
dfe = len(df_europe) + len(df_asia) + len(df_america) + len(df_africa)
- 4

alpha = 0.05
MST = SST / dft
MSE = SSE / dfe
F = MST / MSE
f_critical = stats.f.ppf(1 - alpha, dft, dfe)

print(F, f_critical)
print(F > f_critical)

"""### Since F is larger than the critical value, we have evidence to
reject H0 at alpha = 0.05 and conclude that at least one mean salary of
a continent is different.

### Bonferroni Method
µ1 = The mean salary for individuals in Europe. <br>
µ2 = The mean salary for individuals in Asia. <br>
µ3 = The mean salary for individuals in America. <br>
µ4 = The mean salary for individuals in Africa.
##### H0: µ1 = µ2
##### H1: µ1 = µ3
##### H2: µ1 = µ4
##### H3: µ2 = µ3
##### H4: µ2 = µ4
##### H5: µ3 = µ4
"""

import math
from scipy.stats import t


alpha = 0.05
```

```
n1 = df_europe.shape[0]
n2 = df_asia.shape[0]
n3 = df_america.shape[0]
n4 = df_africa.shape[0]
dfe = n1 + n2 + n3 + n4 - 4

Sp = math.sqrt(MSE)
t_value = t.ppf(1 - alpha / (2 * 6), dfe)

#europe compare to asia
H0_interval_lower = mean_df_europe - mean_df_asia - t_value * Sp *
math.sqrt(1 / n1 + 1 / n2)
H0_interval_upper = mean_df_europe - mean_df_asia + t_value * Sp *
math.sqrt(1 / n1 + 1 / n2)
print(f'H0 interval: {H0_interval_lower} ~ {H0_interval_upper}')
#europe compare to america
H1_interval_lower = mean_df_europe - mean_df_america - t_value * Sp *
math.sqrt(1 / n1 + 1 / n3)
H1_interval_upper = mean_df_europe - mean_df_america + t_value * Sp *
math.sqrt(1 / n1 + 1 / n3)
print(f'H1 interval: {H1_interval_lower} ~ {H1_interval_upper}')
#europe compare to africa
H2_interval_lower = mean_df_europe - mean_df_africa - t_value * Sp *
math.sqrt(1 / n1 + 1 / n4)
H2_interval_upper = mean_df_europe - mean_df_africa + t_value * Sp *
math.sqrt(1 / n1 + 1 / n4)
print(f'H2 interval: {H2_interval_lower} ~ {H2_interval_upper}')
#asia compare to america
H3_interval_lower = mean_df_asia - mean_df_america - t_value * Sp *
math.sqrt(1 / n2 + 1 / n3)
H3_interval_upper = mean_df_asia - mean_df_america + t_value * Sp *
math.sqrt(1 / n2 + 1 / n3)
print(f'H3 interval: {H3_interval_lower} ~ {H3_interval_upper}')
#asia compare to africa
H4_interval_lower = mean_df_asia - mean_df_africa - t_value * Sp *
math.sqrt(1 / n2 + 1 / n4)
H4_interval_upper = mean_df_asia - mean_df_africa + t_value * Sp *
math.sqrt(1 / n2 + 1 / n4)
print(f'H4 interval: {H4_interval_lower} ~ {H4_interval_upper}')
#america compare to africa
H5_interval_lower = mean_df_america - mean_df_africa - t_value * Sp *
math.sqrt(1 / n3 + 1 / n4)
```

```
H5_interval_upper = mean_df_america - mean_df_africa + t_value * Sp *
math.sqrt(1 / n3 + 1 / n4)
print(f'H5 interval: {H5_interval_lower} ~ {H5_interval_upper}')

"""### As a result, we have evidence that µ1 ≠ µ2, µ1 ≠ µ3, µ1 ≠ µ4, µ2
≠ µ3, µ2 ≠ µ4, µ3 ≠ µ4."""
```

7.Appendix

video link: https://youtu.be/G6RFQJwtVMY