Stack Overflow Annual Developer Survey 2023 Results Analysis.

Using the Stack Overflow Annual Developer Survey 2023 data and Python Notebook, we will answer the following questions:

1) How many respondents completed the survey?
2) How many respondents answered all the mandatory questions?
3) What are the median values for the respondents' work experience (WorkExp)?
4) How many respondents work remotely?
5) What percentage of respondents program in Python?
6) How many respondents learned to program through online courses?
7) What is the average and median compensation among respondents who program in Python in each country?
8) What education levels do the top 5 highest paid respondents have?
9) What percentage of respondents program in Python in each age category?
10) Which industries are the most prevalent among respondents (who work remotely) in the 75th percentile for average compensation?

Python

```
import numpy as np
import pandas as pd
survey = pd.read_csv('C:/Users/plish/Desktop/Python_HW/Stack Overflow Developer Survey 2023/survey_results_public.csv')
schema = pd.read_csv('C:/Users/plish/Desktop/Python_HW/Stack Overflow Developer Survey 2023/schema.csv')
```

```
# number of respondents completed the survey
Q1 = survey['ResponseId'].nunique()
print(f'{Q1} respondents took part in the survey')
```

89,184 respondents took part in the survey

```
# number of respondents who answered all the mandatory questions
questions = set(schema.qname.unique()) & set(survey.columns) # connecting survey with survey schema to get all mandatory questions
Q2 = survey.dropna(subset = questions).shape[0]
print(f'{Q2} respondents have answered all the questions in the survey')
```

2,032 respondents have answered all the questions in the survey

```
# median work experience
Q3 = survey['WorkExp'].mode()[0]
print(f'{Q3} years is a value of central tendency for the work experience')
```

5.0 years is a value of central tendency for the work experience

```
# respondents work remotely
remote = survey.loc[survey['RemoteWork'] == 'Remote']
Q4 = remote['ResponseId'].count()
print(f'{Q4} respondents work remotely')
```

30566 respondents work remotely

```
# percentage of respondents who program in Python
respondents = survey['LanguageHaveWorkedWith'].count() # total respondents
python_use = survey.loc[(survey['LanguageHaveWorkedWith'].str.contains('Python', case=False, na=False))]['ResponseId'].count()
Q5 = round((python_use/respondents)*100, 1)
print(f'{Q5}% of respondents program in Python')
```

49.5% of respondents program in Python

```
survey['Python'] = survey['LanguageHaveWorkedWith'].str.contains('Python', case=False,
na=False) # adding new column 'Python' to ease the next tasks solving
```

```
# number of respondents who learned to code via online courses
Q6 = survey.loc[survey['LearnCode'].str.contains('Online Courses', case=False,
na=False)]['ResponseId'].count()
print(f'{Q6} respondents learned to code via Online Courses')
```

43201 respondents learned to code via Online Courses

```
# average and median compensation in each country among respondents who program in Python
clean = survey.dropna(subset='ConvertedCompYearly') # clean data from NaN values
grouped = clean.loc[clean['Python'] == True, ['Country', 'Python', 'ConvertedCompYearly']]
# filtering data for Python users
Q7 = round(grouped.groupby('Country')['ConvertedCompYearly'].agg(['mean', 'median']), 1)
print (f'Average and median compensation among respondents who program in Python in each
country: \n {Q7}')
```

Average and median compensation among respondents who program in Python in each country:

Country	mean	median
Afghanistan	665.0	48.0
Albania	28008.6	11844.0
Algeria	8336.3	6586.0
Andorra	32127.0	32127.0
Angola	662.0	662.0
...	...	...
Venezuela	24973.5	12000.0
Viet Nam	20191.9	13401.0
Yemen	8373.0	9000.0
Zambia	13051.0	9687.0
Zimbabwe	5600.0	6000.0

[158 rows x 2 columns]

```
# education level of top 5 paid respondents
#Q8 = survey.loc[:, ['EdLevel', 'ConvertedCompYearly']].nlargest(5, 'ConvertedCompYearly')
# another way to find it
Q8 = survey.loc[:, ['EdLevel', 'ConvertedCompYearly']].sort_values(by=
'ConvertedCompYearly', ascending=False).head(5)
print('Education levels of Top 5 the highest paid respondents: \n\n', Q8)
```

Education levels of Top 5 the highest paid respondents:

	EdLevel	ConvertedCompYearly
53268	Professional degree (JD, MD, Ph.D, Ed.D, etc.)	74351432.0
77848	Professional degree (JD, MD, Ph.D, Ed.D, etc.)	73607918.0
66223	Bachelor's degree (B.A., B.S., B.Eng., etc.)	72714292.0
28121	Primary/elementary school	57513831.0
19679	Professional degree (JD, MD, Ph.D, Ed.D, etc.)	36573181.0

```
# percentage of respondents who programs in Python in each age category
prog = survey.loc[survey['Python'] == True].groupby('Age')['ResponseId'].count() #
programs in Python
total = survey.groupby('Age')['Python'].count() # total respondents
Q9 = round((prog/total)*100, 2)
# Q9 = survey.groupby('Age').apply(lambda x: round((x['Python'].sum() /
x['Python'].count()) * 100, 2)) # another way using apply and lambda functions
print('Percentage of respondents programming in Python in each age group: \n',
Q9.apply(lambda x : f'{x} %'))
```

Percentage of respondents programming in Python in each age group:

Age	
18-24 years old	61.39 %
25-34 years old	47.58 %
35-44 years old	41.44 %
45-54 years old	38.46 %
55-64 years old	36.50 %
65 years or older	30.91 %
Prefer not to say	41.20 %
Under 18 years old	68.63 %

dtype: object

```
# most prevalent industries among respondents in the 75th percentile for AVG compensation
short = survey.loc[survey['RemoteWork'] == 'Remote', ['ConvertedCompYearly', 'Industry']]
# Remote workers
clean2 = short.dropna(subset='ConvertedCompYearly') # clean salary from NaN values
percent = clean2['ConvertedCompYearly'].quantile(0.75) # 75 percentile

# use mode() aggregation to find the most widespread industries
Q10 = clean2.loc[clean2['ConvertedCompYearly'] == percent]['Industry'].mode()
print(f'Among respondents in the 75th percentile for compensation who work remotely, the most prevalent industries are: \n{Q10}?')
```

Among respondents in the 75th percentile for compensation who work remotely, the most prevalent industries are:

Information Services, IT, Software Development...