Q1) Which among the given options can be used to extract content of the entire page from Wikipedia for the topic “Business Intelligence”. Assume the library is imported using the following command:

```
import wikipedia as wk
```

a) `wk.get_page("Business Intelligence")`

b) `wk.read_page("Business Intelligence")`

c) `wk.page("Business Intelligence")`

d) `wk.save_page("Business Intelligence")`

Q2) Which among the following code blocks will get you the latitude and longitude of “IIT Madras”? Assume the Nominatim library is imported using the command given below:

```
from geopy.geocoders import Nominatim
```

a)
```
location = locator.geocode("IIT Madras, Chennai, India")
print("Latitude = {}, Longitude = {}".format(location.latitude,
location.longitude))
```

b)
```
locator = Nominatim(user_agent="myGeocoder")
location = locator.geocode("IIT Madras, Chennai, India")
print("Latitude = {}, Longitude = {}")
```

c)
```
locator = Nominatim(user_agent="myGeocoder")
location = locator.geocode("IIT Madras, Chennai, India")
print("Latitude = {}, Longitude = {}".format(location.latitude,
location.longitude))
```

d)
```
locator = Nominatim(user_agent="myGeocoder")
print("Latitude = {}, Longitude = {}".format(location.latitude,
location.longitude))
```

Q3) Given below is a function that uses the city name as input and gives us its respective location id as output. This function requires you to provide the location_url for it to work. Among the given options choose the one that helps you get the location_url.

```
def getlocid(city):
    city = city.lower() # convert city name to lowercase to standardize
    format
    # Convert into an API call using URL encoding
    location_url = 'https://locator-service.api.bbci.co.uk/locations?' +
    urlencode({
      'api_key': 'AGbFAKx58hyjQScCXIYrxuEwJh2W2cmv',
      's': city,
      'stack': 'aws',
      'locale': 'en',
      'filter': 'international',
      'place-types': 'settlement,airport,district',
      'order': 'importance',
      'a': 'true',
      'format': 'json'
    })
    result = requests.get(location_url).json()
    locid = result['response']['results']['results'][0]['id']
    return locid
```

a) Right Click on Web Page -> Inspect -> Sources

b) Right Click on Web Page -> Inspect -> Elements

c) Right Click on Web Page -> Inspect -> Performance

d) Right Click on Web Page -> Inspect -> Network

Q4) Provided below is a snippet of the code block of HTML tags from a website providing weather forecast. Your goal is to scrape the high and low values for the 14-day temperature forecast.

```
<div class="wr-day-temperature__high">
        <span class="wr-day-temperature__high-label wr-hide-visually">High</span>
                <span class="wr-day-temperature__high-value">
                <span class="wr-value--temperature ">
                <span class="wr-value--temperature--c">31°</span>
                <span class="wr-hide"> </span>
                <span class="wr-value--temperature--f">87°</span>
                </span>
        </span>
</div>
<div class="wr-day-temperature__low">
        <span class="wr-day-temperature__low-label wr-hide-visually">Low</span>
                <span class="wr-day-temperature__low-value">
                <span class="wr-value--temperature ">
                <span class="wr-value--temperature--c">21°</span>
                <span class="wr-hide"> </span>
                <span class="wr-value--temperature--f">71°</span>
                </span>
        </span>
</div>
```

Also provided below, is the python code to extract values from the tags. But the tags represented as <A> and <B> are missing. Choose the most appropriate tag that will get you the high and low values for the 14-day temperature forecast..

#Daily High Values

```
daily_high_values = soup.find_all('span', attrs={'class': '<B>'})
```

#Daily Low Values

```
daily_low_values  = soup.find_all('span', attrs={'class': '<A>'})
```

a) <A> = wr-value--temperature--f
   <B> = wr-value--temperature--c
b) <A> = wr-value--temperature--c
   <B> = wr-value--temperature--c
c) <A> = wr-day-temperature__low
   <B> = wr-day-temperature__high
d) <A> = low-label wr-hide-visually
   <B> = high-label wr-hide-visually

Q5) The most recent data cannot be obtained when scraping is done using Excel.

a) True
b) False

Q6) Among Scrapy and Beautiful soup libraries, Beautiful soup uses spiders, which are self-contained crawlers that are given a set of instructions to scrape a webpage.

a) True
b) False

Q7) The code block given below prints the CSS/HTML contents of the provided webpage without any errors.

a) True
b) **False**

Code block:

```
#Import necessary libraries
from bs4 import BeautifulSoup as bs
import pandas as pd

#Load the webpage
r = requests.get("https://www.cricbuzz.com/cricket-match/live-scores")
# Convert to a beautiful soup object
```

```
soup = bs(r.content)
# Print out HTML
contents = soup.prettify()
print(contents)
```

Q8) Choose the most suitable pandas module for data preparation activities.

a) Pandas datareader
b) **Pandas profiler**
c) Pandas featuretools

Q9) Pandas profiler provides information about correlation and missing values in the data.

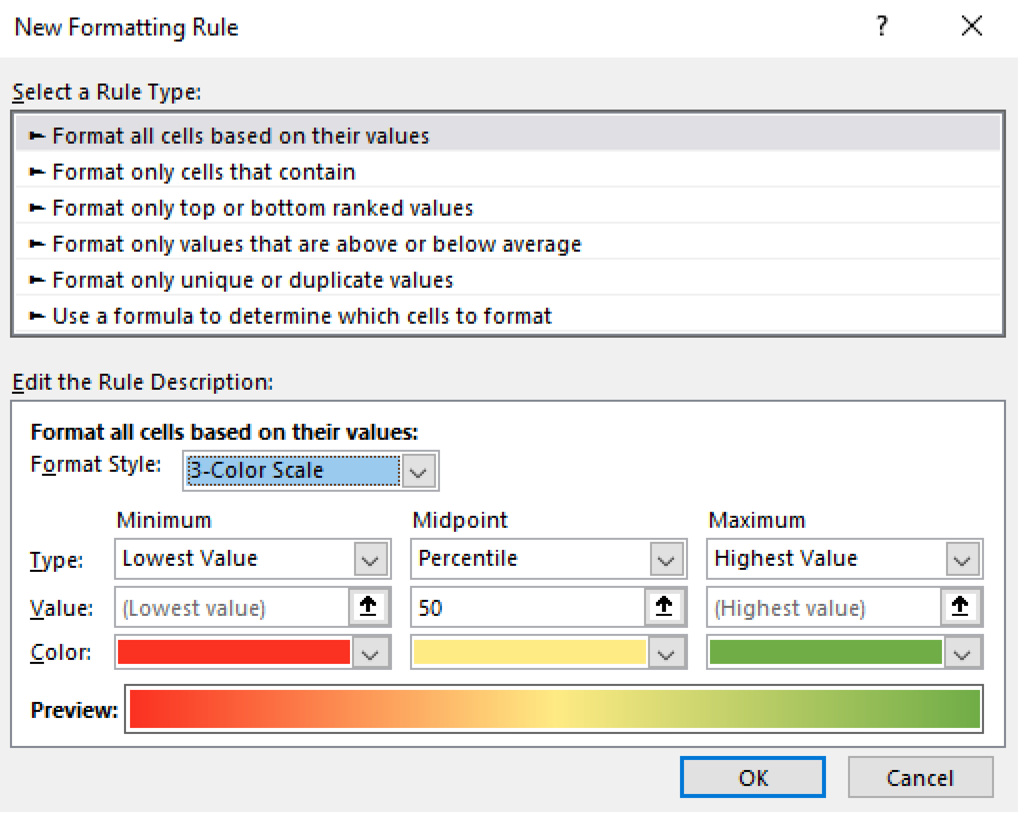
a) True
b) False

Q10) How does converting the data into a table in excel help in data preparation? Choose the most suitable option.

a) It automatically cleans up the null values
b) It highlights the columns with missing values
c) It enables easier referencing of formulae
d) It does not make a difference

Q11) Choose the most appropriate formatting rule based on the scenario given below:

The data consists of cities and their respective population. A city with high population density is undesirable. Your goal is to identify and filter cities that have the highest population density



New Formatting Rule
?
Select a Rule Type:
► Format all cells based on their values
► Format only cells that contain
► Format only top or bottom ranked values
► Format only values that are above or below average
► Format only unique or duplicate values
► Use a formula to determine which cells to format
Edit the Rule Description:
OK
Cancel

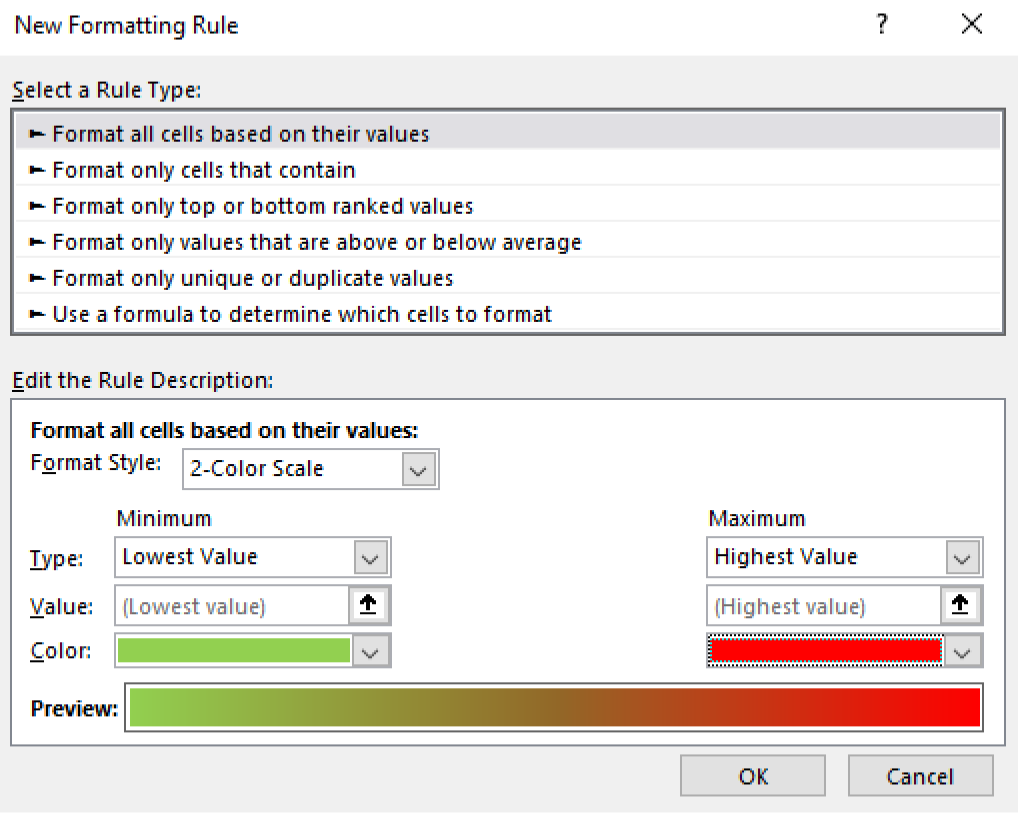
Format all cells based on their values:

Format Style: 3-Color Scale

	Minimum	Midpoint	Maximum
Type:	Lowest Value	Percentile	Highest Value
Value:	(Lowest value)	50	(Highest value)
Color:	🟥	🟨	🟩

Preview:

a)



New Formatting Rule
?
Select a Rule Type:
► Format all cells based on their values
► Format only cells that contain
► Format only top or bottom ranked values
► Format only values that are above or below average
► Format only unique or duplicate values
► Use a formula to determine which cells to format
Edit the Rule Description:
OK
Cancel

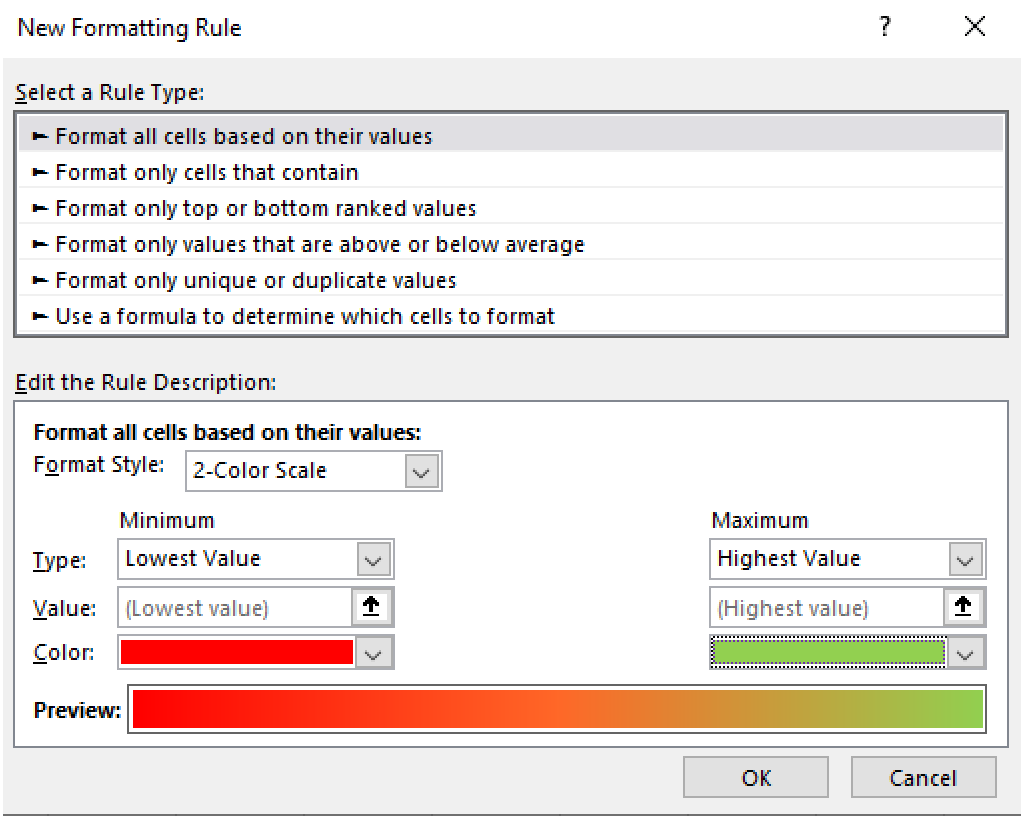
Format all cells based on their values:

Format Style: 2-Color Scale

	Minimum	Maximum
Type:	Lowest Value	Highest Value
Value:	(Lowest value)	(Highest value)
Color:	🟩	🟥

Preview:

b)

c) 

New Formatting Rule

Select a Rule Type:

- Format all cells based on their values
- Format only cells that contain
- Format only top or bottom ranked values
- Format only values that are above or below average
- Format only unique or duplicate values
- Use a formula to determine which cells to format

Edit the Rule Description:

Format all cells based on their values:

Format Style: 2-Color Scale

	Minimum	Maximum
Type:	Lowest Value	Highest Value
Value:	(Lowest value)	(Highest value)
Color:	🟥	🟩

Preview:

Q12) Name an excel feature that helps in data aggregation. Choose the most suitable answer.

a) Azure ML
b) Solver
c) Macros
d) Pivot tables

Q13) Name an excel feature that helps in visualizing trends in data. Choose the most suitable answer.

a) Sparklines
b) Box Plot
c) Color scales
d) Pivot tables

Q14) From the options provided below, identify the steps involved in creating a dataset which consists of images of cricket players from different countries. (Hint: A closely related activity was performed in module 3: Data Preparation)

a) Copy the links of images to a csv file -> Download images of cricket players using Python -> Create a macro to display the images in excel
b) Download images of cricket players using Python -> Copy the links of images to a csv file -> Create a macro to display the images in excel
c) Create a macro to display the images in excel -> Download images of cricket players using Python -> Copy the links of images to a csv file

Q15) Your project requires you to create a dataset consisting of images of cricket players from different countries. Name the windows command that is used to copy the links of images to a csv file. Assume the filename is links.csv.

a) dir /s /b > links.csv
b) dir /b > links.csv
c) dir /s /t > links.csv
d) dir > links.csv

Q16) Open Refine uses clustering algorithms to cluster data. One of the algorithms called key collision uses levenshtein distance to compute the clusters.

a) True
b) False

Q17) Compute the levenshtein distance between the two strings provided below:
String 1: K & M Associates, Inc
String 2: K & J Associates, Ltd

a) 0
b) 2
c) 4
d) 5

Q18) Which of the following set of commands will return links to all pictures in a Wikipedia page of interest as a list?

a)
```
import wikimedia as wk
pg = wk.page("<title of page>")
pg.pictures
```

b)
```
import wikimedia as wk
pg = wk.fullpage("<title of page>")
pg.images
```

c)
```
import wikipedia as wk
pg = wk.page("<title of page>")
pg.images
```

d)
```
import wikimedia as wk
pg = wk.page_url("<title of page>")
pg.images
```

Q19) Our goal is to cluster the days in a month when the sales are high. We use the excel color scale feature on the “Quantity of items Sold” column of the dataset. How do we identify clusters in the data? Choose the most appropriate option.

a) Insert a pivot table to aggregate the data which would provide the clusters

b) Add data bars to identify clusters
c) Zoom out to identify clusters
d) Filter the data in a descending order to identify clusters

Q20) Open Refine uses the following algorithms to cluster similar entities:

a) Key Collision
b) Nearest Neighbor
c) Both a and b
d) None of the above

Q21) A user wishes to scrap a PDF from an url and convert one of the tables in the pdf to a dataframe. Once it is completed, the user wishes to have the descriptive stats and correlation of variables as a report. The order of the libraries which can be used to achieve the same are:

a) Tabula, pandas_profiling
b) Pandas_profiling, Tabula
c) Tabula, cv2
d) None of the above

Q22) Which among the following commands can be used to retrieve titles of content related to “Data Science” in Wikipedia?

a) wk.get_data("Data Science")
b) wk.pull_data("Data Science")
c) wk.search("Data Science")
d) wk.read_pages("Data Science")