Introduction

Dear prospect employee -

Welcome to RSR and we are pleased to have you in our data engineering team. As a pre-boarding step, please complete the reading task and the small project in Python assigned to you. This is like a refresher task that will prepare you for your job and helps to better understand the RSR structure in a better way and will ease your onboarding.

Reading Task:

1. Self learning Skills:

Explore gcp documentation and explain the below:

cloud storage bucket
- Definition and concept
- retention period
- multi regional storages?

Compute Engine
- Definition and concept
- What are instances ?
- Instance Types
- Storage Types

Bigquery
- Definition and concept
- What are datasets ?

Kubernetes
- Definition and concept
- Please go through this video

https://cloudacademy.com/course/introduction-to-google-kubernetes-engine-gke/cluster-architecture/

Outcome/expectation: after you join the team, you need to prepare a ppt slideshow about your understanding of the above cited topics and present it to the team.

reviews Python and OOPs concepts:

Read the below example and understand the concepts clearly.

https://www.programiz.com/python-programming/object-oriented-programming

Implement the requirement cited below similar to the above example.

==========
processing a ETL on data
==========

In your case, extraction will be a .csv file from your local system - a folder: "source_data"
and Loading will be into another folder: "dest_data"

---

use this dataset page:
https://www.kaggle.com/datasets/rajanand/education-in-india?select=2015_16_Statewise_Elementary.csv

Extract three three datasets from here:
statewise_elementary
statewise_secondary
district_wise

Please download them as .csv file into your local system - in a folder: "source_data"

==========
MAIN TASK SUMMARY
==========

task 1:

identify the key columns in each dataset.

task 2:

identify the common column in all 3 datasets

task 3:

Use the common column as a key and combine the datasets.

So there should be one large dataset, that is a combination of elementary and secondary datasets and district wise

task 4:

calculate percentage_urban_population
group data on district, and give the number of sch1 schools in each district in Jammu Kashmir,

==========
MAIN TASK DETAILS
==========

create a project folder and name it ast: "indian_school_data"
create two modules within this project as explained below.

module#1 : extract data

create a module: "mod_extract_data".py
and a class "cls_extract_data"
create a method to read the csv files and return a dataframe.
So you should pass a file path and the method should return a dataframe.

module#2: process data

create another module. call it "mod_process_data.py"
in this "mod_process_data.py":

Base class creation in module 2

create a base class / parent class. "base_process_data"
add common methods inside it. Think of some common methods you can add.
Please add a couple of them of your own, as you see fit. A few such ideas / thoughts for you to think through, for instance:

fetch_data
    must fetch data using the "mod_extract_data".py module.
get_shape
clean_data_for_nulls
select_imp_columns

Child classes creation in module 2

you will add 2 child classes:
one for elementary, and another for secondary.

- "process_data_elem"
- "process_data_secd"

inherit the base class in the child classes.
implement the parent class methods in the child class.
*** pass the appropriate parameters and return the appropriate info

methods creation in child class

add the below new methods in child class "process_data_elem":

1. "func_combine_datasets"

create three subset datasets by taking a few important columns only from each dataset. Choose wisely.
Must include any key columns, TOTPOPULAT, P_URB_POP, SCH1 ; tot_population, urban_population, sch1; columns from both datasets.
now create one big dataset from these three datasets by joining them on their key columns.
and return this big dataset.

1a. "func_calculate_urban"
calculate_percentage of urban population using the above combined_dataset and return a new dataset with new column called "percent_urban"
output this dataset as "ds_urban_percent".csv in "dest_data" folder

1b. "func_regional_dataset"
calculate the count of number of sch1 schools in in each district in Jammu Kashmir.
output this dataset as "ds_district_schools_jk".csv in "dest_data" folder

In your computations to use inheritance, a sample link is provided and you can use example #3 as your guide for inheritance from this example.
https://www.programiz.com/python-programming/object-oriented-programming

module#3: main module

Finally create a main module "main.py"
And create a class called "process_main"
and 1 method called "run"

run the entire app from this main.py
So that means, you have to efficiently import modules to execute them all in order.