# BDT Jr Data Analyst - “Homework” Instructions

Thanks for interviewing with Benefits Data Trust! In order to gauge your skills as a data analyst, we would like you to prepare a brief analysis for us review. We have outlined an analysis below.

* The data is provided in a public dataset in Google BigQuery. You can set up a free account. If you run into difficulties with this, let us know and we can help – you will not be assessed negatively if you need help.
* Provide the code you use and visualizations you create pasted in the document below or in a zip file.

## Assessment Criteria

You will be assessed on your problem-solving, programming & analysis work, and clarity of presentation. This work will be an initial investigation/exploration of the data. It does not need to be exhaustive but can simply cover the questions below.

To be practical as an interview exercise, if you are uncertain about anything, make a reasonable assumption and note that. You will be assessed on your analytical work, not on your knowledge of the benefits domain.

If you have questions about the interview process, please reach out to your recruiter.

## Data

Within Google BigQuery (<https://console.cloud.google.com/bigquery>) there is a dataset called `bigquery-public-data` which contains the table `census\_bureau\_acs.county\_2018\_5yr`. This table contains census data grouped at the county level for all counties in the US. All counties in PA have a geo\_id that starts with the digits “42”.

## Questions

Please answer the question and include any code you used to find the answer. When providing a visualization, use any tool that you like.

#Query used to extract data

SELECT \*

 FROM `bigquery-public-data.census\_bureau\_acs.county\_2018\_5yr`

 WHERE geo\_id LIKE '42%';

1. According to this data, what is the total population of PA? What do you observe about how population is distributed across counties in PA?   
    **- The total population of PA: 12,791,181.0**

**- There are three major counties that make up over 28% of the population and the majority of counties are rather small**

1. What is the average poverty rate in PA? Poverty rate is the number of people in poverty over the total population.  
    **-The average poverty rate in PA is 12%**
2. How many Pennsylvanians live in counties where the poverty is higher than the state average?

**- The number of Pennsylvanian's living in counties where poverty is above the state average: 5,418,448.0**

1. Is the poverty rate in PA consistent across all counties or is there variation? Provide a visualization to explain your answer.

-**There appears to be relatively consistent poverty rates with only one major outlier that also happens to be PA's most populated county. There is a low standard deviation of .03 which means that most of the data is clustered around the mean and relatively consistent.**

1. Does there seem to be any relationship between a county’s poverty rate and the percent of population that is over 65? Include a visualization in your response.  
    **- I do not see a relationship between a county's population over 65 and their poverty rate. Population over 65 had a weaker correlation (.0016) than that of total population and poverty rate (.0788)**
2. How many households are on public assistance or food stamps in PA?  
    **-Number of households on public assistance/food stamps in PA: 700,759.0**

**-Percent of households on public assistance/food stamps in PA: %13.672614664310846**

1. In looking at the county level data for poverty and for public assistance/food stamps, are there any counties where there might be a large ‘gap’, meaning that county has more people in poverty who have not enrolled in benefits, compared to other counties? What assumptions do you need to make to answer this question?

*(NOTE: Generally people qualify for public assistance or food stamps at or below 185% of the poverty line. )*

**-Assuming everyone within the 'Poverty' column qualifies for assistance this is the results. County #42027 has the lowest usage rate by far of 16.5%. From there it is a gradual upslope to the highest usage rate of 60.9% from county #42073. It is clear that some counties are not maximizing the possible assistance that is available to them.**

I have attached a copy to my file I used in Python as well as a link to my file on GitHub.

<https://github.com/lvdonofrio14/PA-County-Census-Analysis>