Tools for Data Analytics

Student Submission Form

To complete this assessment, you need to create a ZIP archive folder that includes all the files and code used complete the data analysis organized in three subfolders named "Part A", "Park B," and "Part C," as they related to the part of the ask, and a completed copy of this form in the main folder. You will upload your zipped folder that includes this completed form and the subfolders to Taskstream to complete your submission. Use as many rows in the tables below as necessary and remove those not used.

Student Name: Desiree Teter

WGU Student ID: #000809943

Inventory of Part A-related files (Subfolder "Part A"):

Version of Python used: 3.6

Python Libraries used (if any): csv, BeautifulSoup, SoupStrainer, requests, urllib.request, urlopen, re, urllib.parse, urljoin

Platform Python was used on: Sublime + Windows Command Prompt

Name of the PDF file with the responses to task prompts from Part A:
PartAWriteUp.PDF

Name(s) and description(s) of Python files:

Name	Extension	Description
classscraper	**py**	**Script to scrape unique HTML links and output to csv**

Name(s) and description(s) of input file(s) from Part A:

Name	**Extension**	**Description**
https://www.census.gov/programssurveys/popest.html / viewsource_https___www.census.gov_programssurveys_popest	html	HTML file/web page scraped by python script.

Name(s) and description(s) of output file(s) from Part A:

Name	**Extension**	**Description**
scriptrun	**PNG**	**Screenshot of Python Script completed on Command Line**
externallinks	**CSV**	**Output of Python script**

Inventory of Part B-related files (Subfolder "Part B"):

SQL Environment used: MySQL

Platform SQL was used on: MySQL Workbench 6.3 CE

Name of the PDF file with the responses to task prompts from Part B: PartBWriteUp.PDF

Name(s) and description(s) of SQL code files:

Name	**Extension**	**Description**
scriptfirsttask	**sql**	Script to create table with absolute mathematical differences between estimates for two years (Task I)

scriptsecondtask	**sql**	Script to create table with differences between estimates of 10000 or more between two datasets, rounded to 100s (Task J)
scriptthirdtask	**sql**	Script to create table with differences between estimates of 10000 or more between two datasets, rounded to 10000s, with year columns (Task L)

Name(s) and description(s) of input file(s) from Part B:

Name	**Extension**	**Description**
2016	csv	Cleaned data for 2016
2017	csv	Cleaned data for 2017
nst-est2016-01.xlsx	xlsx	Raw data for 2016
nst-est2017-01.xlsx	xlsx	Raw data for 2017

Name(s) and description(s) of output file(s) from Part B:

Name	**Extension**	**Description**
absdifffinal	**csv**	CSV contaiing output for task L table of differences between estiamtes > 10000 rounded to 10000s
absdiff	**csv**	CSV containing output for Task J table of differences between estimates > 10000, rounded to 100s
tabletwo	**csv**	CSV containing output for task I table of absolute differences between two years estimates

Inventory of Part C-related files (Subfolder "Part C"):

R version used: `R version 3.4.2 (2017-09-28)`

R packages used (if any): bigmemory, psych, pastecs, dplyr, reshape2

Platform R was used on: RStudio

Name of the PDF file with the responses to task prompts from Part C: PartCWriteUp.pdf

Name(s) and description(s) of R script(s):

Name	**Extension**	**Description**
script	**R**	**Script containing all assignment tasks, comments added**

Name(s) and description(s) of input file(s) from Part B:

Name	**Extension**	**Description**
statesestimates2	.csv	Cleaned data
nst-est2017-01.xlsx	xlsx	Raw data

Name(s) and description(s) of output file(s) from Part C:

Name	**Extension**	**Description**
naturalhist	**png**	Histogram for task O
fivehundredthouhist	**png**	Histogram for task O
regressionline	**png**	Plotted regression line for linear model for recent years population estimates, Task Q
estimatestatdescription	**png**	Statistical summary of results Task P

2020prediction	**png**	Plotted regression line for predictions for 2020 based Task M