Data Science Case Study - Analytics

At Nielsen we work with a wide variety of datasets from different sources. We would like to see how you approach working with a new dataset to perform data analysis and provide meaningful insights into the data.

Scope and Expectations:

- We are interested in seeing how you use data and code to *answer questions* and *derive insights* into data trends. We know your time is limited. You will have several days to complete this task, but please do not spend more than **4 hours total time** on it. Should you be asked to continue through the interview process, you may be asked to discuss this problem further.
- Please use either python or SQL as primary languages when coding your work.
- We do not expect you to have perfect answers. We want to see your thought process for how you derived the answers.
- We would like to see all of your code, results, and documentation. This includes your data analysis, visualization, evaluation, written comments, etc.
- Deliverables will be evaluated based on the following competencies: **Communication, Critical Thinking, Data Cleaning and Transformation, Data Analysis & Visualization, and Coding Proficiency.**

Deliverables

Please email the following items to your recruiter or team member by the date specified:

1. **Documentation** - A summary of your findings and answers to the questions provided in both parts. You should also include a brief description of what you did and why, including details of the tools you used to derive your data insights (i.e., data visualization tools such as tableau, etc.).
2. **Code** - The code you used to explore the data and apply statistical techniques. Please include all data cleaning or data transformation steps you may have performed as part of this exercise. Please document your code so that the reviewer can easily follow how the code is being used.

You may combine items 1 and 2 into a single document or notebook, but keep in mind the reviewer must be able to review any output.

If you have questions, experience issues opening the dataset, or will not be able to complete the task in the time given, please contact your recruiter or team member as soon as possible.

Part One: Population Statistics

The data for this exercise comes from the U.S. Census Bureau. It contains population statistics for counties in the U.S. from 2010 to 2015. Given several years of historical data, **your goal is to understand how the population has changed over time by county and state and whether there are any changes observed in demographic trends.**

Data structure and contents:

All of the data is contained in the county_census_population file. Each row of data contains information for a particular county and year. The year, county name, and state in which the county is located are also given.

Additional columns show the total female population (`female_total_population`) and male population (`male_total_population`), and various age groups ranging from under 5 years old to over 85 years old. For example, the `female_age_30_to_34` shows the count of females that are 30 to 34 years old. The `male_age_10_to_14` shows the count of males that are 10 to 14 years old. *Tip:* You may need to perform data cleaning or data transformation steps to ensure the data is usable.

Analysis:

Please perform the following tasks and provide your answers to the following questions.

1. Import and clean the data set.
2. Please provide a table of summary statistics for the total population for each state (not county).
3. Choose 5 counties to analyze further. These should be representative of the country.
    a. Describe how/ why you chose these 5 counties to analyze.
    b. Please provide visualizations to depict population trends over time.
    c. Have there been significant shifts in demographic trends for any of these 5 counties across the 6 years? If so, please describe.
4. If you were to continue analyzing all of the counties provided in this data set, what would your next steps be?

Part Two: Top 10 Best Cities to Live In

The data for this exercise comes from CNN Money's online list of the Best Places to Live for the years of 2012 and 2013. This list ranks the top small towns to live in across America during that year.

Data structure and contents:

All of the data is contained in the top_ten_small_cities_US file. Each row of data contains information for a particular city and year. The county name and state in which the county is located are also given. Additional columns show the rank the city was given on CNN's list, and the year it was given that rank. You will also find the population for the city for that year, as provided by CNN.

Analysis:

Please perform the following tasks and provide your answers to the following questions in the same document as provided in Part One.

1. Import and clean the data set (if needed).
2. Merge this data set with the US Population data from part one. The resulting data set should contain information for only the counties mentioned in the top_ten_small_cities_US file, but should contain data for all 6 years (2010 through 2015).
3. What percent of the total population for each county resides in each of the cities on the list? (for example, what percent of Hamilton County resides in Carmel, IN?)
4. Are there any shared population characteristics of these counties that made the Top Ten list? Please explain.
5. Please provide at least one visualization of Fairfax County, Virginia. This should include an indication of the year(s) any of the cities in this county made it in CNN's Top Ten Cities list.