Data Insight

Knowledge for Insight from Data

DATA SCIENTIST PROGRAM

Project: Time Series Analysis of NAICS

A Data Analyst is someone who munges information using data analysis tools. The meaningful results they pull from raw data help their employers or clients make important decisions by identifying various facts and trends. The North American Industry Classification System (NAICS) is an industry classification system developed by the statistical agencies of Canada, Mexico, and the United States. NAICS is designed to provide common definitions of the industrial structure of the three countries and a common statistical framework to facilitate the analysis of the three economies.

For this task, consider yourself a Data Analyst, which you definitely are after months of data skill training facilitated by Data Insight. The head of your department at your new company has given you the following instructions:

1. Download zipped files A_NEWLY_HIRED_DATA_ANALYST.zip (**PASSCODE>2021DSP**)

2. Prepare a data set using the following files:

a. NAICS 2017 – Statistics Canada: Description of the North American Industry Classification System (NAICS). All you would need to understand for this task is, how the NAICS works as a hierarchical structure for defining industries at different levels of aggregation. For example (see page 22), a 2-digit NAICS industry (e.g., 23 - Construction) is composed of some 3-digit NAICS industries (236 - Construction of buildings, 237 - Heavy and civil engineering construction, and a few more 3-digit NAICS industries).

Similarly, a 3-digit NAICS industry (e.g., 236 - Construction of buildings), is composed of 4-digit NAICS industries (2361 - Residential building construction and 2362 - Non-residential building construction).

b. Raw data: 15 CSV files beginning with RTRA. These files contain employment data by industry at different levels of aggregation; 2-digit NAICS, 3-digit NAICS, and 4-digit NAICS. Columns mean as follows:

(i) SYEAR: Survey Year
(ii) SMTH: Survey Month
(iii) NAICS: Industry name and associated NAICS code in the bracket
(iv) _EMPLOYMENT_: Employment

c. **LMO Detailed Industries by NAICS:** An excel file for mapping the RTRA data to the desired data. The first column of this file has a list of 59 industries that are frequently used. The second column has their NAICS definitions. Using these NAICS definitions and RTRA data, you would create a monthly employment data series from 1997 to 2018 for these 59 industries.

d. **Data Output Template:** An excel file with an empty column for employment. You should fill the empty column with the data you prepared from your analysis.

e. **Take note of the following:** (i) The industry names in the 'LMO Detailed Industries by NAICS' match with the industry names in the 'Data Output Template'. The RTRA data should be used based on the NAICS codes, not by industry names.

(ii) Try to create each series from the highest possible level of aggregation in the raw data files. For example, if an LMO Detailed Industry is defined with a 2-digit NAICS only, do not use a lower level of aggregation (i.e., 3-digit or 4-digit level NAICS files in the RTRA). Similarly, if an LMO Detailed Industry is defined with a 3-digit NAICS only, do not use the 4-digit NAICS files for that industry.

(iii) All steps, including merging or appending the data, that would generate the requested data should be done using python codes.

(iv) The source for the data is: Real Time Remote Access (RTRA) data from the Labour Force Survey (LFS) by Statistics Canada.

3. **Highlight at least 5 important questions** that you would like to answer in order to provide valuable information to your company so that they can make good business decisions. Your questions should include how employment in Construction evolved over time and how this compares to the total employment across all industries? **and at least 4 more questions**.

4. **Carry out a detailed time series** and any other data analysis to answer your questions above. Your analysis should be done in the notebook of your choice (e.g., Jupiter, Colab, Deepnote, etc.) with sections, headings, and comments that aid readability.

5. Use visualizations to illustrate each answer. Visualizations could be a graph, a combination of graphs, or a dashboard. Add any necessary details to them, as if these visualizations would be published or presented to external stakeholders.

6. **Submit your complete notebook** to your GitHub repository and provide the link here.

7. **Write a blog** about your new role as a Data Analyst, **NAICS,** the analysis you performed, your findings, and recommendations for your company and those who are aspiring to be a Data Analyst. Give your blog a unique title. The blog should be written here (https://www.datainsightonline.com/blog) with code from your notebook embedded (**Do not embed code images!** See the options you have for embedding your code).

The completion of this project is due on January 19, 2022.