DSC520 Week-10 Exercise 10.3

Venkat Jagadeesh Jampani

February 20th 2022

Project: Impact of Covid 19 on Home Prices/IT Jobs/Migration in USA. – Step:2

How to import and clean my data?

Import Data:

Identify the data source, in the current project, we have different sources of data: 1. Covid cases,and covid deaths reported across different states. 2. Jobs Data : number of jobs openings across different states in USA. 3. Hosing prices and its raise from previous year across different states in USA.

Data cleaning:

Data cleaning is the process of ensuring data is correct, consistent and usable. You can clean data by identifying errors or corruptions, correcting or deleting them, or manually processing data as needed to prevent the same errors from occurring.

Most aspects of data cleaning can be done through the use of software tools, but a portion of it must be done manually. Although this can make data cleaning an overwhelming task, it is an essential part of managing company data.

Benefits of data cleaning? There are many benefits to having clean data:

1. It removes major errors and inconsistencies that are inevitable when multiple sources of data are being pulled into one dataset.
2. Using tools to clean up data will make everyone on your team more efficient as you'll be able to quickly get what you need from the data available to you.
3. Fewer errors means happier customers and fewer frustrated employees.
4. It allows you to map different data functions, and better understand what your data is intended to do, and learn where it is coming from.

Data cleaning in six steps The first step before starting a data cleaning project is to first look at the big picture. Ask yourself: What are your goals and expectations?

To achieve those goals you've set, next, you must plan a data cleanup strategy. A great guideline is to focus on your top metrics. Some questions to ask:

1. What is your highest metric looking to achieve?
2. What is your company's overall goal and what is each member looking to achieve from it? A good way to start is to get the key stakeholders together and brainstorm.

3. Monitor errors Keep a record of trends where most of your errors are coming from.This will make it a lot easier to identify and fix incorrect or corrupt data. Records are especially important if you are integrating other solutions with your fleet management software, so that your errors don't clog up the work of other departments.

4. Standardize your process Standardize the point of entry to help reduce the risk of duplication.

5. Validate data accuracy Once you have cleaned your existing database, validate the accuracy of your data. Research and invest in data tools that allow you to clean your data in real-time. Some tools even use AI or machine learning to better test for accuracy.

6. Scrub for duplicate data Identify duplicates to help save time when analyzing data. Repeated data can be avoided by researching and investing in different data cleaning tools that can analyze raw data in bulk and automate the process for you.

7. Analyze your data After your data has been standardized, validated and scrubbed for duplicates, use third-party sources to append it. Reliable third-party sources can capture information directly from first-party sites, then clean and compile the data to provide more complete information for business intelligence and analytics.

8. Communicate with your team Share the new standardized cleaning process with your team to promote adoption of the new protocol. Now that you've scrubbed down your data, it's important to keep it clean. Keeping your team in the loop will help you develop and strengthen customer segmentation and send more targeted information to customers and prospects.

Final Data Set :

```
1. Covid-Information.csv
2. Housing data-USA.csv
3. Jobs-data.csv
```

Can you reduce your data by selecting only certain variables?

```
1. In all the data sources, select the appropriate and related data only. and filter out and delete the
2. Delete Alpha numeric values and special values formm the data.
```

Could creating new variables add new insights?

```
1. New variables like home price and decrease in each state from the previous year.
2. Increase or decrease percentage of housing data from the previous year.
3. Increase or decrease in number of jobs in each state compare to previous year.
4. Percentage of increase or decrease of number of jobs in each state from the previous year.
```

Questions for Future step :

1. What features could you filter on?
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
4. Could creating new variables add new insights?
5. Could summary statistics at different categorical levels tell you more?
6. How can you incorporate the pipe (%>%) operator to make your code more efficient?