**C2 Lessons Learned**

Below are key coding in Python that was helpful. By validating a dataset has been cleaned before any correlation or processing of the data occurs, any analytical tools will be accurate without erroneous data.

1. Learned how to connect to an SQL database
   1. db\_connection\_str = 'mysql+pymysql://deepanalytics:Sqltask1234!@34.73.222.197/deepanalytics'
2. Learned how data can be imported with different data types when imported into Jupyter via SQL or CSV file.
3. How to look for null values and clean up
   1. dfc.isnull().any()
   2. dfc[dfc['MyUnknownColumn'].isnull()]
4. How to drop rows and columns
   1. dfc\_r1.drop('Unnamed: 0',axis=1, inplace=True)
5. How to remove duplicate rows and opt-in for keeping the first or last row of duplication
   1. df.loc[df.duplicated(keep='last'),:]
6. pandas\_profiling.ProfileReport is extremely useful. It exponentially speeds up data validation process and helps validate errors in data.
7. When using get\_dummies, there is are biases that’s created for correlations that need to be taken into account.

In the future, I’d recommend first running pandas\_profiling.ProfileReport then trying to clean up data. The report is extremely usefully for initial investigation and validating data has been cleaned after updates applied.