**Description of data preprocessing:**

* Data Cleansing:
  + Raw data has a data quality index, the team decided to use data evaluated as Excellent, Very good, Good and Acceptable to have enough data to serve the project problem and questions.
  + Raw data has missing values for years 2020 and 2021 due to Covid-19. The team decided to drop these values as they may skew our machine learning models output.
  + 2015 Q1 has much fewer data points compared to the remaining quarters.
  + Raw data contains three types of statistics, Job Vacancies, Proportion of Job Vacancies, Average Offered Hourly Wage, Machine learning Model will use Job Vacancies statistics as primary dataset. Other datasets will be used to enrich visualization and data interpretation.

* Data Transformation:
  + Several columns will need transformation, text extraction and encoding to be usable in the machine learning model i.e. NOC, Job Characteristics, all types, full time and parttime will all be encoded to be used in our machine learning model.
  + Unnecessary columns will be dropped to preserve space and simplicity, these columns will be: DGUID, UOM/UOMID since we are focusing on Job vacancies it is assumed to be a Number value, SCALAR\_FACTOR, SCALAR\_ID, VECTOR, COORDINATE, STATUS, SYMBOL, TERMINATED, DECIMALS.

**Description of feature engineering and the feature selection including the team's decision-making process**

The team executed initial data exploration to review the current features available in the data set. Then evaluate what features can be most relevant to the project’s problem

.

The team proposed to have the following features as input:

·   **Reference Date:** quarterly periods, in format of string which we will convert to int for analysis.

·   **National Occupancy Classification:** The national classification of jobs (NOC), this is also a string and will be encoded using one hot encoding method

·   **Job Characteristics:** different Job characteristics and requirements (full time vs part time, education level required) this is also a string and will be encoded using one hot encoding method

·   **Previous Job Vacancies:** total number of vacancies available in correlations to reference data, stored as an integer.

·   **Geographical Location (Geo)**: the province in which the vacancies exist.

The output feature will be:

**Current Job vacancies:** current total number of vacancies available, stored as an integer.

**Description of how data was split into training and testing sets**

The data will be split into training and testing randomly with stratification. However, there will be specific requirements for the split data, if the feature is built for a specific date it must consist of previous data i.e feature built for Jan 2016 must consist of Oct 2015 data.

**Explanation of model choice plan**

The project problem and desired output calls for a supervised machine learning model. The team will evaluate regression models using sample data and decide on the best model that fits the project’s purpose. This is because we are treating the data as continuous, using this method we will predict the trend of job vacancies in Canada in different National Occupational Classifications.

**Technology**

The current dataset we utilize for our machine learning model has been cleaned once and currently stored on google drive at 500mb. We will use Google Collab to run our machine learning module.