Project Assignment 1: Data Wrangling

Credible

Menghong Han

Peihan Tian

Yanghe Liu

Laurence Finch

Wrangling with Renewables

Introduction

The state of California has the goal of 100% “zero-carbon” sourced power by 2045. A preliminary target is for 60% of its power to be from renewable sources by 2030. Our data set contains hourly data on the amount of energy produced from various sources of renewable energy including solar.

The business relevance of renewable energy sources is obvious. Given California’s, and other states’/countries’ emissions reduction targets there is growing demand for renewables and there are opportunities for high growth.

Climate change in general, driven by greenhouse gas emissions partly from fossil fuel powered energy companies is a very pressing issue for businesses as well as society more broadly. The implications of climate change are far-reaching and will affect all industries from agriculture to energy as environmental conditions worse, consumer awareness and preferences change and environmental disasters become more frequent. These challenges will pose a threat to some businesses, and offer opportunities to others.

Data source

The data set is from Kaggle https://www.kaggle.com/cheedcheed/california-renewable-production-20102018

The original source is California ISO (CISO): an independent non-profit, public-benefit corporation power grid operator, who supply 80% of California's power lines. The nature of the CISO and their transparent and open approach to their data makes us believe that this is a highly reliable source of data.

An initial inspection finds that there are some missing values, especially in the solar variable.

CISO states that the data is unverified raw data and is not intended to be used as the basis for operational or financial decisions, so we knew from the start that it would be necessary to clean it.

all_breakdown

TIMESTAMP	BIOGAS	BIOMASS	GEOTHERMAL	Hour	SMALL HYDRO	SOLAR	SOLAR PV	SOLAR THERMAL	WIND TOTAL
2011-07-21 16:00:00	175.0	375.0	961.0	17.0	577.0	412.0			1109.0
2011-07-21 17:00:00	176.0	369.0	963.0	18.0	589.0	363.0			1322.0
2011-07-21 18:00:00	177.0	367.0	965.0	19.0	553.0	303.0			1539.0
2011-07-21 19:00:00	177.0	366.0	968.0	20.0	553.0	259.0			1624.0
2011-07-21 20:00:00	179.0	364.0	970.0	21.0	554.0	273.0			1707.0
2011-07-21 21:00:00	177.0	365.0	973.0	22.0	536.0	274.0			1775.0
2011-07-21 22:00:00	177.0	347.0	976.0	23.0	522.0	179.0			1815.0
2011-07-21 23:00:00	176.0	336.0	977.0	24.0	523.0	0.0			1769.0
2013-05-29 00:00:00	210.0	326.0	895.0	1.0	343.0		0.0	0.0	3454.0

Figure 1 Original Data

Data cleaning

As we went through the data, the first step we did was to check the extreme values of each variable. Since the dataset has hourly reports on the power production from various sources, all numerical variables should contain positive numbers. However, we found out there's one negative number in variable WIND TOTAL, so we replace it with 0.

Then, as we found out variable SOLAR missed around 70 percent of data, we chose to delete the whole column. Also, for the missing values in SOLAR PV and SOLAR THERMAL, we saw there's a pattern in it. Data of SOLAR PV and SOLAR THERMAL are missing at the same time, and all missing values combine to be in a whole missing day. So, the missing data of solar energy seems only due to random missing. Therefore, we just deleted all missing rows of SOLAR PV and SOLAR THERMAL.

As a result, since all missing values are missing completely at random, we believe our data is relatively clean.

TIMESTAMP	BIOGAS	BIOMASS	GEOTHERMAL	Hour	SMALL HYDRO	SOLAR	SOLAR PV	SOLAR THERMAL	WIND TOTAL
Apr 2010 - Jan 2018	0 - 248	0 - 455	0 - 1,230	1 - 24	500 - 520 1544 2.28%	0 - 1,069	0 - 9,373	0 - 725	-21 - 4,914
2011-07-21 00:00:00	177.0	372.0	984.0	1.0	516.0	0.0			1607.0
2011-07-21 01:00:00	176.0	373.0	984.0	2.0	509.0	0.0			1619.0
2011-07-21 02:00:00	177.0	371.0	985.0	3.0	506.0	0.0			1660.0
2011-07-21 03:00:00	177.0	374.0	985.0	4.0	504.0	0.0			1633.0
2011-07-21 04:00:00	177.0	377.0	984.0	5.0	508.0	0.0			1520.0
2011-07-21 05:00:00	177.0	371.0	980.0	6.0	510.0	0.0			1406.0
2011-07-21 06:00:00	176.0	372.0	979.0	7.0	521.0	10.0			1391.0
2011-07-21 07:00:00	176.0	378.0	977.0	8.0	524.0	84.0			1341.0
2011-07-21 08:00:00	176.0	380.0	975.0	9.0	526.0	169.0			1220.0
2011-07-21 09:00:00	175.0	381.0	973.0	10.0	528.0	254.0			1025.0
2011-07-21 10:00:00	175.0	383.0	971.0	11.0	530.0	339.0			798.0
2011-07-21 11:00:00	175.0	384.0	969.0	12.0	532.0	418.0			711.0
2011-07-21 12:00:00	173.0	386.0	966.0	13.0	563.0	451.0			703.0
2011-07-21 13:00:00	174.0	387.0	965.0	14.0	573.0	450.0			761.0
2011-07-21 14:00:00	174.0	380.0	964.0	15.0	575.0	446.0			853.0
2011-07-21 15:00:00	175.0	370.0	963.0	16.0	570.0	435.0			977.0
2011-07-21 16:00:00	175.0	375.0	961.0	17.0	577.0	412.0			1109.0
2011-07-21 17:00:00	176.0	369.0	963.0	18.0	589.0	363.0			1322.0
2011-07-21 18:00:00	177.0	367.0	965.0	19.0	553.0	303.0			1539.0
2011-07-21 19:00:00	177.0	366.0	968.0	20.0	553.0	259.0			1624.0
2011-07-21 20:00:00	179.0	364.0	970.0	21.0	554.0	273.0			1707.0
2011-07-21 21:00:00	177.0	365.0	973.0	22.0	536.0	274.0			1775.0
2011-07-21 22:00:00	177.0	347.0	976.0	23.0	522.0	179.0			1815.0
2011-07-21 23:00:00	176.0	336.0	977.0	24.0	523.0	0.0			1769.0
2013-05-29 00:00:00	210.0	326.0	895.0	1.0	343.0		0.0	0.0	3454.0
2013-05-29 01:00:00	210.0	323.0	895.0	2.0	341.0		0.0	0.0	3268.0
2013-05-29 02:00:00	210.0	326.0	896.0	3.0	342.0		0.0	0.0	3202.0
2013-05-29 03:00:00	210.0	322.0	896.0	4.0	342.0		0.0	0.0	3007.0
2013-05-29 04:00:00	209.0	313.0	891.0	5.0	345.0		2.0	0.0	2563.0
2013-05-29 05:00:00	208.0	308.0	840.0	6.0	352.0		4.0	0.0	2290.0
2013-05-29 06:00:00	208.0	309.0	839.0	7.0	351.0		106.0	0.0	2829.0
2013-05-29 07:00:00	200.0	312.0	839.0	8.0	353.0		436.0	34.0	2801.0
2013-05-29 08:00:00	202.0	308.0	836.0	9.0	357.0		857.0	220.0	2715.0
2013-05-29 09:00:00	200.0	314.0	785.0	10.0	361.0		1202.0	308.0	2461.0
2013-05-29 10:00:00	199.0	325.0	805.0	11.0	360.0		1442.0	316.0	2350.0

Figure 2 Dirty Dataset

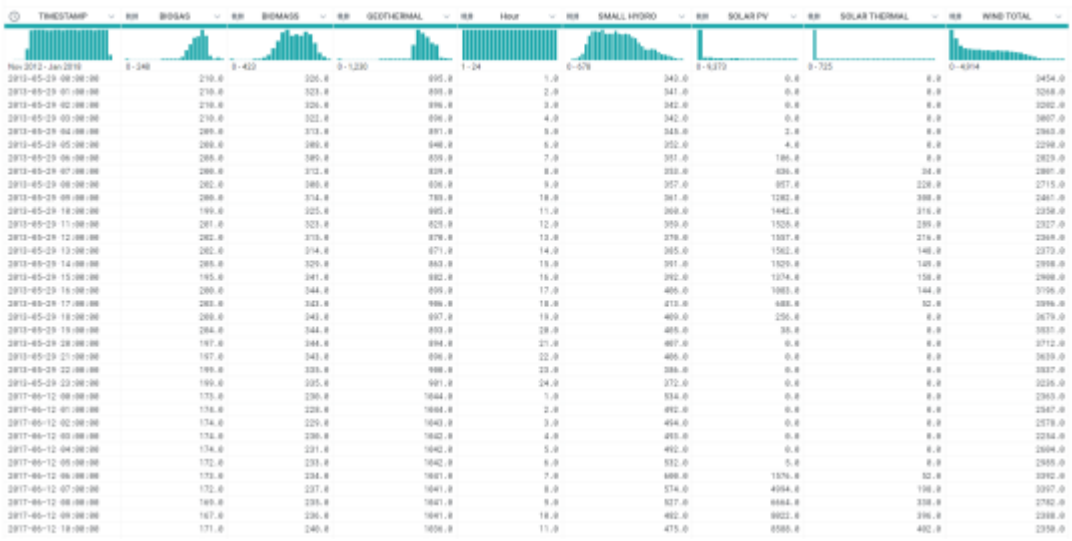

Figure 3 Cleaning Dataset

Descriptive statistics and data quality

After data cleaning, all our 9 variables have 100% valid data which means no mismatched or missing values.

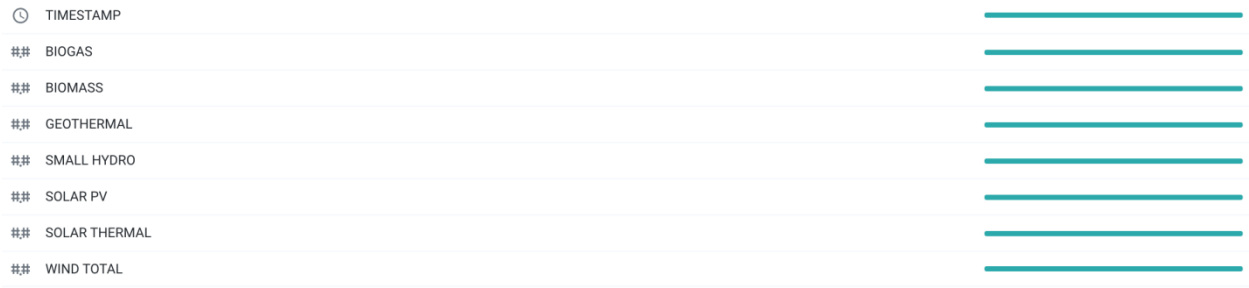

Figure 4 Variables in Dataset

Then, we choose the following 6 key variables to demonstrate data quality: "BIOGAS" , "BIOMASS" , "GEOTHERMAL", "SMALL.HYDRO", "SOLAR.PV", "WIND.TOTAL".

Note: We found the distribution of "SOLAR.PV" and "SOLAR.THERMAL" are quite similar during data cleaning, therefore, we choose "SOLAR.PV" to show data quality.

First of all, we draw histograms with normal density distribution for each variable to measure the distribution of data.

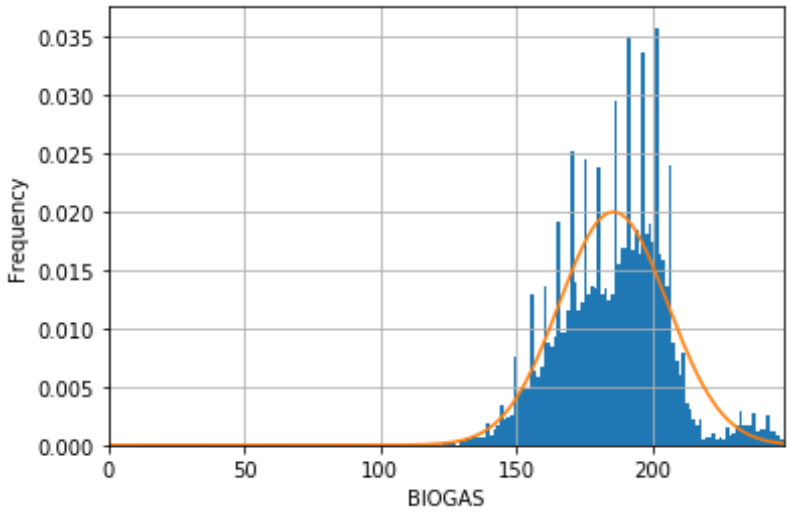

Figure 5 Distribution of Biogas

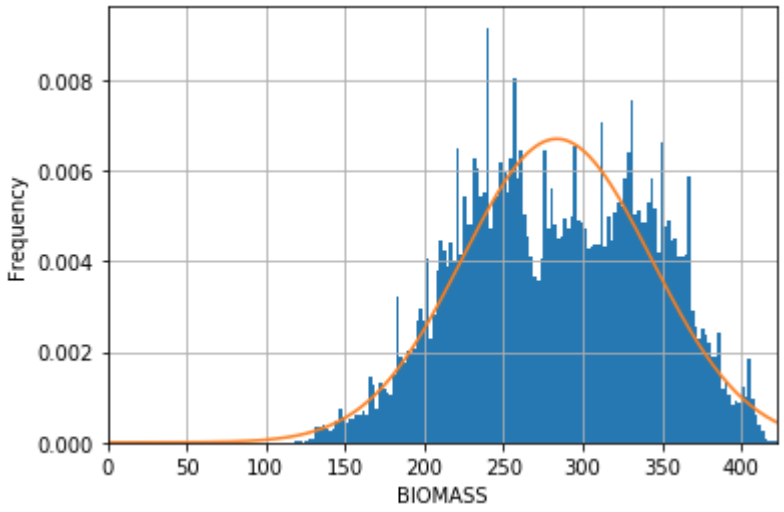

Figure 6 Distribution of Biomass

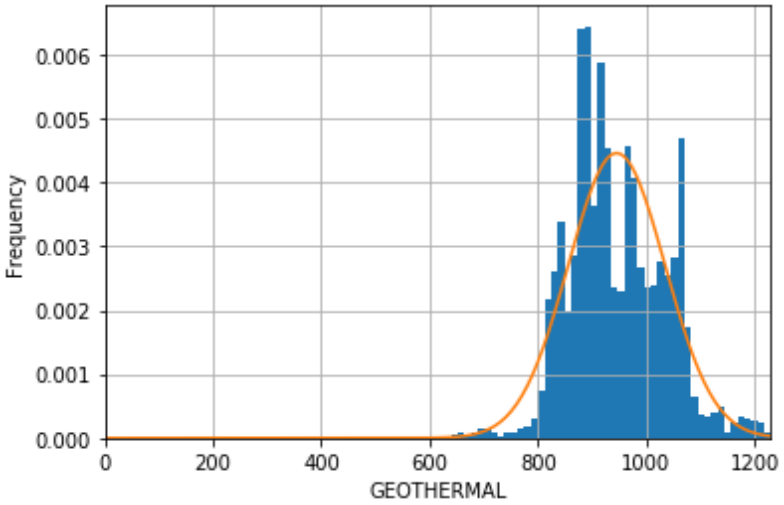

Figure 7 Distribution of Geothermal

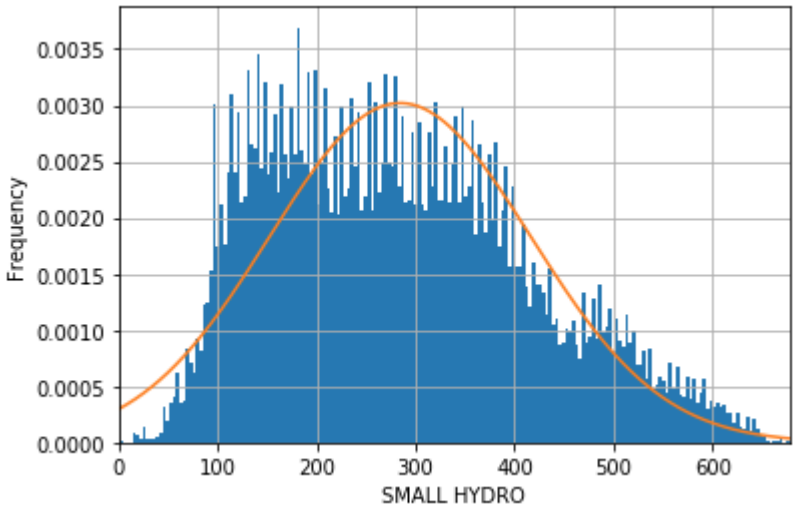

Figure 8 Distribution of Small Hydro

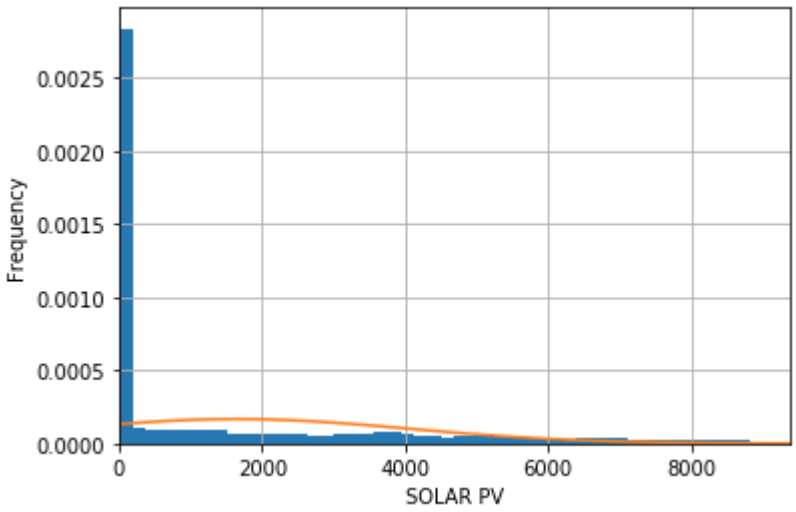

Figure 9 Distribution of Solar PV

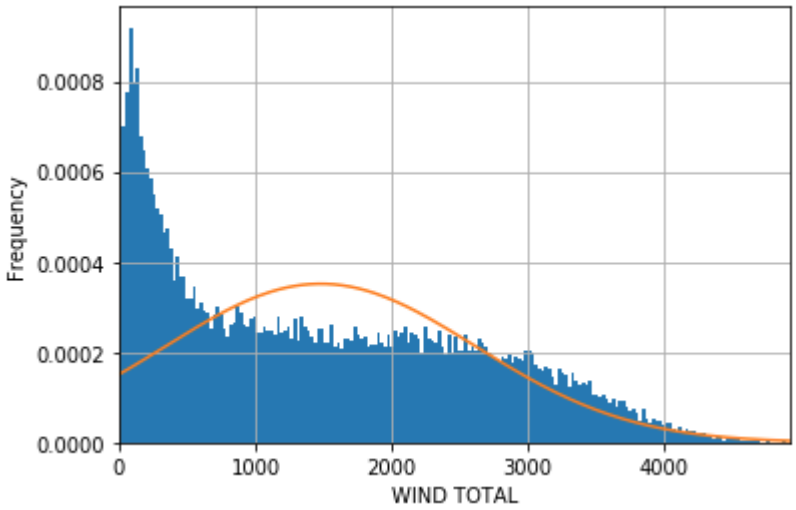

Figure 10 Distribution of Wind Total

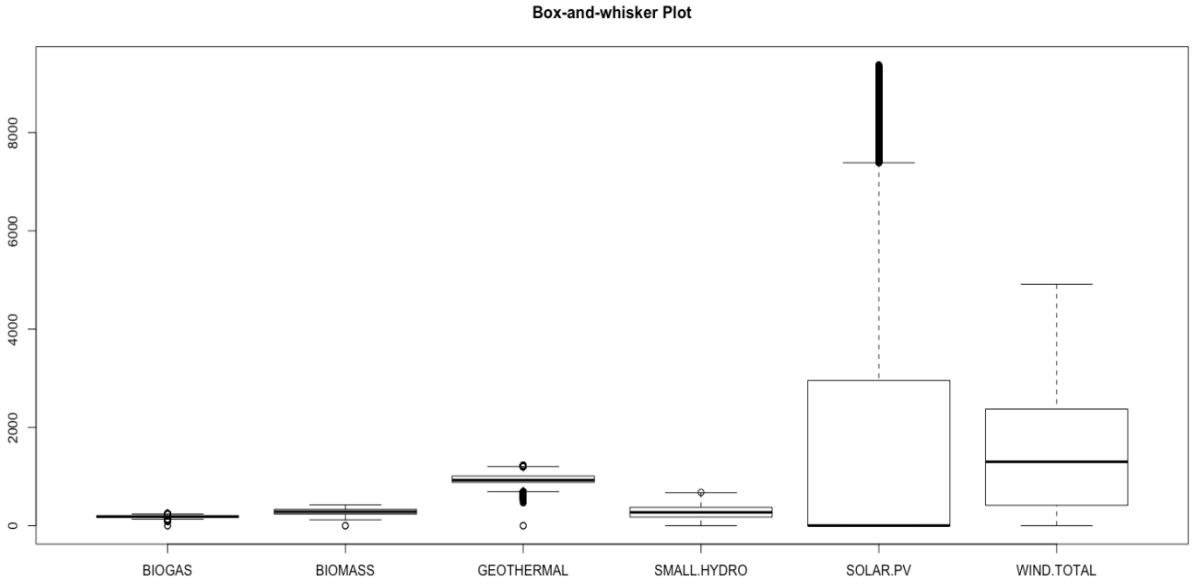

Figure 11 Boxplots of Variables

From the distribution chart, we can see the distribution of the first 4 variables, "BIOGAS", "BIOMASS", "GEOTHERMAL", "SMALL.HYDRO" are relatively close to normal distribution compared to the distribution of the last two variables "SOLAR.PV" and "WIND.TOTAL". Let's look at the first 4 variables in detail.

From the statistic summary, the distribution of "BIOGAS" presents a slightly negative skewness and a kurtosis slightly larger than 3 (skewness = -0.03, kurtosis= 0.81), "BIOMASS" shows negative skewness and a kurtosis slightly less than 3 (skewness = -0.06, kurtosis= -0.78),

"SMALL.HYDRO" shows a large positive skewness and a kurtosis slightly less than 3(skewness = 0.45, kurtosis= -0.52), "GEOTHERMAL" shows a slightly positive skewness and a kurtosis larger than 3 (skewness = 0.14, kurtosis= 0.92). The parameters show the distributions of the above variables are rather close to normal distribution. Furthermore, based on the box plots of those 4 variables below we can see fewer outliers therefore we conclude that the data quality of the first 4 variables seems good in general.

From the distribution of "SOLAR.PV", we found that data highly concentrated on the frequency of 0 and shows positive skewness. However, this makes sense since "SOLAR.PV" is Solar Photovoltaic which works only when the sun rises. From the box chat of "SOLAR.PV" we can see the outliers show a large right deviation that is to say although Solar Photovoltaic works relatively fewer hours a day, it generates much more energy than others in a short period of time. Therefore, the large number of outliers would not destroy the data quality of "SOLAR.PV".

The distribution of "WIND.TOTAL" is similar. From evening to morning next day and during the noon, there would not be so much wind which explains the positive skewness of the distribution. From the box chat of "WIND.TOTAL", we can also see right deviation of outliers which illustrate the comparatively high intensity of wind power during a short period of time. Therefore, the data quality of "WIND.TOTAL" seems good.

Summary

A, in your analytical modeling processing;

Before building models, we have to ensure the quality of the data, which means the absence of missing values, outliers, the presence of standardized or normalized data. What kinds of methods we use to clean data always depends on what kind of models we want; causality, or prediction models. For example, if we want to build a regression model for prediction and we only have several variables and millions of rows. In this kind of situation, we can add interaction terms to our models to decrease its MSE. We can also standardize the data to reduce errors caused by units. We should also check the distribution of our variables, which must be normal distribution because it will fit the assumption of OLS.
To fulfill this goal, we can use Trifacta Wrangler, which can automatically find missing values, outliers and standardize our data easily. With the help of it, we will definitely ensure data quality and prepare better for the next step: modeling.

B, in combination with other software tools;

From data manipulation, we found Trifacta Wrangler works well in combination with other softwares. First of all, it's very convenient to import and export data in the forms of cvs or excel for data preparation and we can use Python, R, SQL or Tableau to do further data mining. In addition, Trifacta Wrangler has partnership with AWS, Google cloud, Microsoft Azure and Snowflake, and this enable users to accelerate data cleaning and preparation with modern platforms for cloud data lakes & warehouses as well as ensure the success of your analytics, ML & data onboarding initiatives across any cloud, hybrid and multi-cloud environment. Take AWS as an example, Trifacta Wrangler provides an Intelligent Data Prep Solution for the Cloud on AWS. As an AWS certified ML Competency and Data & Analytics Competency partner, Trifacta offers an enterprise-class data preparation solution that natively integrates with an expansive set of AWS services and AWS database services including Amazon S3, Amazon EMR, Amazon Redshift, Amazon SageMaker and Amazon IAM. Whether users are migrating to a cloud data lake with Amazon S3, modernizing the legacy data warehouse to Amazon Redshift, or launching your ML/AI project in Amazon SageMaker, users can rely on Trifacta's industry-leading data preparation solution to fuel mission-critical cloud projects with clean, connected and timely data at all times.

C, in a collaborative workflow within or across organizational boundaries;
We can use Trifacta Wrangler to create flow and share our result of wrangled data within or across organization boundaries. Through the shared flow, people are able to see the modification on the datasets. Also, the recipe part in Trifacta Wrangler helps people in a group know which process data cleaning had gone through. Another useful part of Trifacta Wrangler is that it allows us to create new flows, which can let us make modification on the dataset without influence the other flows. As a result, Trifacta Wrangler makes collaborative workflow more effective and less hazardous to accidents.