The following text is taken from the capstone project report chapter on Data Wrangling

2. DATA WRANGLING

A good amount of data wrangling and cleansing were required to prepare the data for analysis. The primary data issues found were: 1) typographical errors; 2) missing values; 3) duplicated records in the production spreadsheets; 4) differences in unique well identifier (API) formatting and difficulties associating the correct sidetrack numbers (API with 10 digits) in the production records with the parent API8 (API with 8 digits) in the surface shapefile; 6) N-COM production assignments in the first few months of a well production which needed reassignment to the correct producing formation; and 5) other inconsistencies between the surface, bottomhole, and production datasets. Many of these issues required a first pass at cleansing, merging, and comparison to identify, so the process was an iterative one. The following section describes the steps taken during the data wrangling and cleansing process.

Reading data into Pandas DataFrames

With all but a few exceptions, the production report files were formatted and named similarly, so the 'glob' module of python could be used for batch loading using Jupyter Notebook. The 22 separate files were then appended together into one large pandas dataframe, and a separate list was loaded to define the column labels. The resulting raw dataframe (allproddf) consisted of 12,922,083 rows and had a file size of over 3.2 GBytes.

The three shapefile datasets and an area of interest polygon were loaded into a separate Jupyter notebook using the geopandas module. Each of the three resulting geo-dataframes were then clipped to the area of interest polygon. The procedure was to iterate through the x, y coordinates (geometry) for each well. If the well's coordinates fell within the area of interest polygon, it was added to a shortlisted API well list. This list was then used to filter those wells into new subsetted geo-dataframes. These three new filtered geo-dataframes were also written out to new filtered shapefiles for later merging with the production information. The WELL.SHP file resulted in the gdf_surf_aoi dataframe. The DIRECTIONAL_BOTTOMHOLE_LOCATIONS.SHP file resulted in the gdf_bh_aoi dataframe. And finally the DIRECTIONAL_LINES.SHP file resulted in the gdf_dirlines dataframe.

An important feature contained in the gdf_surf_aoi dataframe was the ground elevations of the wells. Many of these values were missing and were required to calculate the correct bottomhole subsea bottomhole depth of the wellbore. This required an additional iteration step (See the Missing Values section below).

Consistent API Formatting

The API number is the unique well identifier and a key attribute shared between datasets. However, the format of the attribute varied between the datasets, so reformatting was required. For example, the API was parsed into separate attributes of 'api_county_code', 'api_seq_num', and 'sidetrack_num' in the production dataset and these attributes were defined as integers. In the shapefiles, the API was an object padded with leading zeros. In the WELLS shapefile, the value was limited to 8 digits, while the other shapefiles included two additional digits (sidetrack number) appended on the end. The API attribute was converted to integers with the leading zeros in the county code removed to match the production dataframe. The 8 digit API in the WELLS shapefile was renamed to API8, and two API attributes were created in the production dataframe, an API column and an API8 column.

Production Data Redundancy

There was some redundancy between the separate yearly production reports which resulted in duplicate rows for the same API and date, once merged. These were removed by subsetting on the date, API, and formation_code attributes and dropping the duplicates. The last row was kept with the assumption that the earlier records appended to the file first are superceded by corrected information from a later date. This brought the resulting dataset down to 11,910,725 rows. The formation_code attributes needed some subsequent re-formatting to remove variants, so the de-duplication process was run a second time after to remove additional duplicates due to these formation_code variants (See Categorical Variables section below).

Handling Datetime

Separate month and year attributes in the production dataset needed to be converted to a datetime series. These separate attributes were converted to strings and the month attribute was padded with a leading zero. Initial qc of the month column indicated one row with an erroneous month value of '0', which was preventing the conversion to a datetime series from completing successfully. This row (116971) was for a well outside the study area boundary, so was dropped. Then the 'Date' attribute could be successfully created to a datetime series and the dataframe indexed using this series. End-month???

The Stat_Date and Spud_Date attributes in the WELLs shapefile also needed conversion to datetime. The replacement of two erroneous Stat_Date values of '3019-12-19' and '2029-09-06' had to be corrected to '2019-12-19' and '2019-09-06' respectively. Erroneous Spud_Date values with years of '2029', '2109' and '2108' were also corrected to 2019 and 2018 respectively in 6 wells.

In addition the prod_days attribute was converted to a timedelta dtype for direct comparison with a calculated feature 'total elapsed production time'.

Categorical Variables

The formation_code and well_status attributes are important categorical identifiers and had varying formats which had to be corrected. Lowercase and mixed case strings were all converted to uppercase. Trailing spaces in the formation_codes were removed. This brought the number of unique well_status categories down from 15 to 9 and the number of unique formation_code categories down from 155 to 100. Both fields were converted from objects to categoricals.

It was noted during analysis of the formation_code attribute, that the code 'N-COM' is assigned to early production in many of the wells, prior to an actual formation assignment. To get more accurate production volumes by formation_code, the N-COM code was reassigned to the actual formation ('fm_code_realloc') where known. Other wells having a formation_code of N-COM throughout their production history were left intact.

Missing Data

Approximately 3000 wells were missing important Ground_Ele information in the surface location shapefile. This attribute is crucial to calculating the bottomhole depth (TVDSS) values for the wellbores. The missing ground elevations were imputed using the Shapely nearest_points, Point, and MultiPoint functions for Python. The location of the nearest wells with a ground_elevation was taken to impute the value.

Additionally, some monthly production records for producing wells were missing. These were imputed using a rolling moving average value from the available monthly production data for those wells.

Outliers

A few outliers were identified in the TVD attribute max and min ranges, which needed correction to calculate accurate TVDSS values. Five wells were found with extraordinarily large values of 73699, 77018, 82090, and 70993250. Fortunately, the correct values could be ascertained by querying the COGCC's COGIS on-line database. Examination of the correct TVD values for those wells indicated that they should be clipped at the first 4 digits. One extraordinarily small value of -4553 was corrected to a positive value after confirmation with the online COGIS database. Other outliers were values of 0,7,150,217, and 558 in directional wells with MDs of over 7000 ft and were revised to the MD of the well.

There are far too many vertical wells with Max_MD = 0 to correct properly. These were excluded from theTVDSS calculation.

19 horizontal wells were mistakenly identified as well_type_cat of 'Vertical'. These were identified by their extraordinarily large Max_MD values. They were re-categorized to well_type_cat' = 'Horizontal'.

44 wells had an incorrect Facil_Stat which did not correspond to the production records and well_status assignment from the production records. For instance, a 'DA' well (drilled and abandoned) that had known production through to the end of the production records was corrected to 'PR'. And other wells assigned as 'DG' (Drilling), 'XX' (Permitted Location), and 'AL' (Abandoned Location) with current known production were also corrected to 'PR'.

38 rows in the production dataframe had prod_days greater than 31 days, which is impossible for one month. These were reset to the maximum number of days in the month.

Monthly reported water volumes for many wells were extraordinarily and unrealistically large. Many of these wells were actually injection wells in which the injected water volume has been systematically entered into the same water volume fields as produced volumes. For this reason, injection wells were removed from the dataset (183 vertical and directional wells). Other unusually large monthly water volumes for 10 producing wells were found to be data entry errors and were replaced with values consistent with the production in that well from past and future months.

Merging DataSets

The production time-series dataframe was merged with the gdf_surf_aoi geodataframe using an inner join to produce a datetime dataframe subsetted to only producing wells within the AOI boundary.

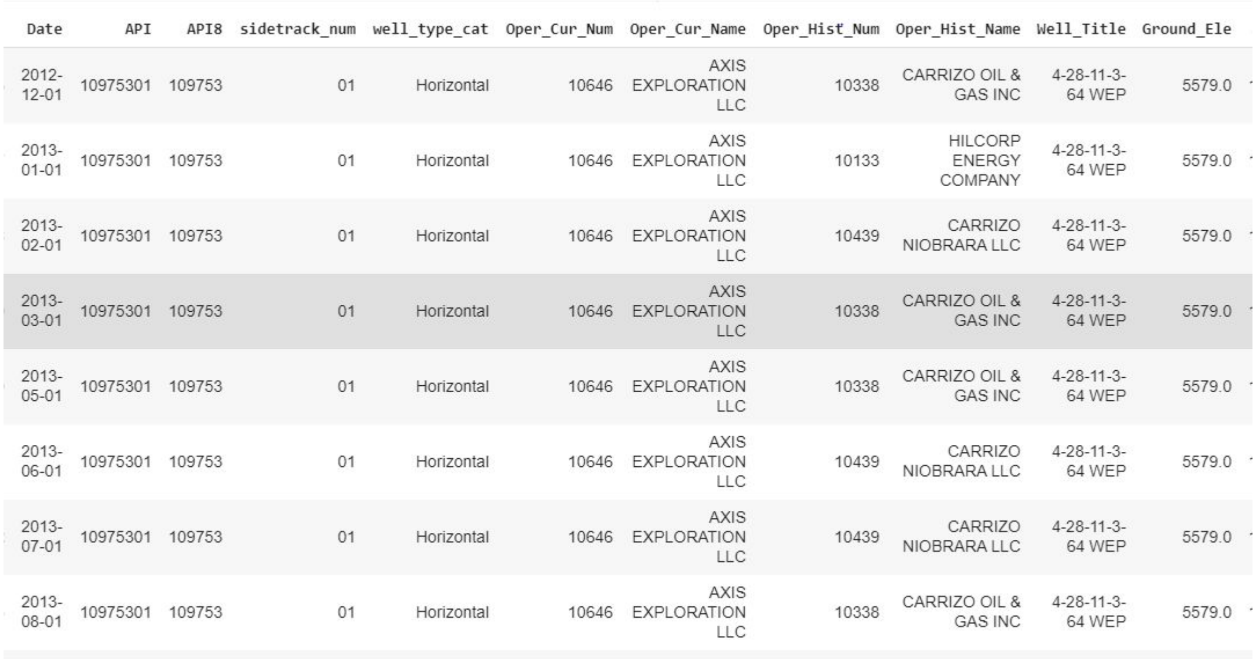
The dataframe was also joined with the bottomhole location geodataframe (gdf_bh_aoi) to obtain the bottomhole locations and other pertinent attributes of the deviated and horizontal boreholes. This join was performed in a left fashion to maintain the complete list of producing wells.

Removing Unwanted Columns and Renaming Others

Unnecessary columns were dropped and the remaining columns were re-organized to bring the most pertinent and key attributes to the left side of the dataframes. Additional feature attributes were calculated that were pertinent. These features will be discussed in the Feature Engineering section of this report.

Final Production TimeSeries DataFrame

The final production time-series dataframe (Prod_DT_Series_Final_WQuantileRank) has a total of 6,393,783 rows and 82 columns and is over 3.5 GBytes in size. A slice of the dataframe is shown below. A full list of the attributes is provided in section 9.5 of the Appendix.



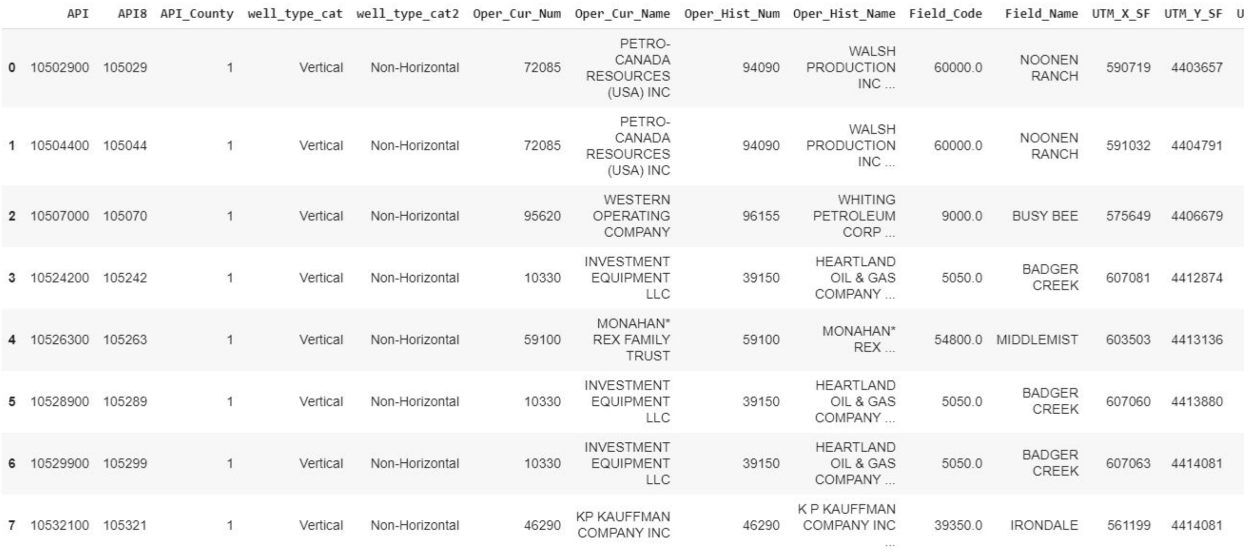
Date	API	API8	sidetrack_num	well_type_cat	Oper_Cur_Num	Oper_Cur_Name	Oper_Hist_Num	Oper_Hist_Name	Well_Title	Ground_Ele
2012-12-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10338	CARRIZO OIL & GAS INC	4-28-11-3-64 WEP	5579.0
2013-01-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10133	HILCORP ENERGY COMPANY	4-28-11-3-64 WEP	5579.0
2013-02-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10439	CARRIZO NIOBRARA LLC	4-28-11-3-64 WEP	5579.0
2013-03-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10338	CARRIZO OIL & GAS INC	4-28-11-3-64 WEP	5579.0
2013-05-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10338	CARRIZO OIL & GAS INC	4-28-11-3-64 WEP	5579.0
2013-06-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10439	CARRIZO NIOBRARA LLC	4-28-11-3-64 WEP	5579.0
2013-07-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10439	CARRIZO NIOBRARA LLC	4-28-11-3-64 WEP	5579.0
2013-08-01	10975301	109753	01	Horizontal	10646	AXIS EXPLORATION LLC	10338	CARRIZO OIL & GAS INC	4-28-11-3-64 WEP	5579.0

Fig 3: Slice of final production time-series dataframe

Final Rollup DataFrame

To gather statistics for each well and perform further exploratory analysis on non-time-series features associated with each well, a rollup dataframe was created. The final rollup dataframe (rollup_prodhead_final) has 49,280 rows (the number of unique wells and producing formation pairs in the final production time-series dataframe) and 66 columns and is 21.8 MBytes in size. A slice of the final rollup dataframe is shown below.

See the Feature Engineering section for more discussion on the attributes within this dataframe. A full list of the attributes is provided in section 9.6 of the Appendix.



	API	API8	API_County	well_type_cat	well_type_cat2	Oper_Cur_Num	Oper_Cur_Name	Oper_Hist_Num	Oper_Hist_Name	Field_Code	Field_Name	UTM_X_SF	UTM_Y_SF	U
0	10502900	105029	1	Vertical	Non-Horizontal	72085	PETRO-CANADA RESOURCES (USA) INC	94090	WALSH PRODUCTION INC ...	60000.0	NOONEN RANCH	590719	4403657	
1	10504400	105044	1	Vertical	Non-Horizontal	72085	PETRO-CANADA RESOURCES (USA) INC	94090	WALSH PRODUCTION INC ...	60000.0	NOONEN RANCH	591032	4404791	
2	10507000	105070	1	Vertical	Non-Horizontal	95620	WESTERN OPERATING COMPANY	96155	WHITING PETROLEUM CORP ...	9000.0	BUSY BEE	575649	4406679	
3	10524200	105242	1	Vertical	Non-Horizontal	10330	INVESTMENT EQUIPMENT LLC	39150	HEARTLAND OIL & GAS COMPANY ...	5050.0	BADGER CREEK	607081	4412874	
4	10526300	105263	1	Vertical	Non-Horizontal	59100	MONAHAN* REX FAMILY TRUST	59100	MONAHAN* REX ...	54800.0	MIDDLEMIST	603503	4413136	
5	10528900	105289	1	Vertical	Non-Horizontal	10330	INVESTMENT EQUIPMENT LLC	39150	HEARTLAND OIL & GAS COMPANY ...	5050.0	BADGER CREEK	607060	4413880	
6	10529900	105299	1	Vertical	Non-Horizontal	10330	INVESTMENT EQUIPMENT LLC	39150	HEARTLAND OIL & GAS COMPANY ...	5050.0	BADGER CREEK	607063	4414081	
7	10532100	105321	1	Vertical	Non-Horizontal	46290	KP KAUFFMAN COMPANY INC	46290	K P KAUFFMAN COMPANY INC ...	39350.0	IRONDALE	561199	4414081	

Fig 4: Slice of final rollup-datafram

Related Appendix Materials

9.5 List of Attributes Contained within the Final Production Time-Series Dataframe

Column Index	Attribute	Dtype
0	Date	datetime64[ns]
1	API	int64
2	API8	int64
3	sidetrack_num	object
4	well_type_cat	category
5	Oper_Cur_Num	int64
6	Oper_Cur_Name	object
7	Oper_Hist_Num	int64
8	Oper_Hist_Name	object
9	Well_Title	object
10	Ground_Ele	float64
11	Max_MD	float64
12	MD	float64
13	Max_TVD	float64
14	TVD	float64
15	TVDSS	float64
16	Field_Code	float64
17	Field_Name	object
18	Spud_Date	datetime64[ns]
19	Stat_Date	datetime64[ns]
20	well_status	category
21	Facil_Stat	category

22	API_Form	object
23	formation_code	category
24	fm_code_realloc	object
25	prod_days	float64
26	water_vol	float64
27	oil_vol	float64
28	gas_prod	float64
29	gas_prod_boe	float64
30	LAT_SF	float64
31	LONG_SF	float64
32	LAT_BH	float64
33	LONG_BH	float64
34	UTM_X_SF	int64
35	UTM_Y_SF	int64
36	UTM_X_BH	float64
37	UTM_Y_BH	float64
38	Township	object
39	Range	object
40	Section	object
41	water_disp_code	object
42	water_press_tbg	float64
43	water_press_csg	float64
44	bom_invent	float64
45	adjustment	float64
46	eom_invent	float64
47	gravity_sale	float64
48	gas_vol	float64
49	shrink	float64
50	gas_press_tbg	float64
51	gas_press_csg	float64
52	facility_name	object
53	facility_num	object
54	accepted_date	object
55	revised	object
56	year	object
57	month	object
58	api_seq_num	object
59	API_Label_x	object
60	Well_Num	object
61	Well_Name	object
62	Citing_Typ	object

63	Facil_Id	int64
64	Facil_Type	object
65	Loc_Qual	object
66	Loc_ID	float64
67	Loc_Name	object
68	Dist_N_S	float64
69	Dir_N_S	object
70	Dist_E_W	float64
71	Dir_E_W	object
72	Qtr_Qtr	object
73	Meridian	object
74	BH_Status	object
75	geometry_SF	geometry
76	geometry_BH	geometry
77	ProdHist	category
78	NBRR_Hor_IP_Quintile	category
79	NBRR_Hor_NormBoeCum_Quintile	category
80	CODL_Hor_IP_Quintile	category
81	CODL_Hor_NormBoeCum_Quintile	category
82	prod_month_by_API_Form	int64

9.6 List of Attributes Contained within the Final Rollup Dataframe

Column Index	Attribute	Non-Null Count	Dtype
0		49280	int64
1	API8	49280	int64
2	API_County	49280	int64
3	well_type_cat	49280	category
4	well_type_cat2	49280	category
5	Oper_Cur_Num	49280	int64
6	Oper_Cur_Name	49280	category
7	Oper_Hist_Num	49280	int64
8	Oper_Hist_Name	49280	category
9	Field_Code	49280	float64
10	Field_Name	49280	category
11	UTM_X_SF	49280	int64
12	UTM_Y_SF	49280	int64
13	UTM_X_BH	22269	float64
14	UTM_Y_BH	22269	float64

15	well_status	49280	category
16	Facil_Stat	49280	category
17	Stat_Date	49269	datetime64[ns]
18	Ab_Val	48277	float64
19	TVDSS	48734	float64
20	Spud_Date	44969	datetime64[ns]
21	API_Form	47158	object
22	fm_code_realloc	47158	object
23	Start	47158	datetime64[ns]
24	End	47158	datetime64[ns]
25	oil_cum	47158	float64
26	gas_boe_cum	47158	float64
27	gas_mcf_cum	47158	float64
28	wtr_cum	47158	float64
29	boe_cum	47158	float64
30	prod_days	47158	timedelta64[ns]
31	norm_oil_cum	47158	float64
32	norm_wtr_cum	47158	float64
33	norm_gas_mcf_cum	47158	float64
34	norm_gas_boe_cum	47158	float64
35	norm_boe_cum	47158	float64
36	gor	42532	float64
37	wor	42532	float64
38	30Day_Oil	47158	float64
39	30Day_GasBoe	47158	float64
40	30Day_ProdDays	47158	float64
41	30Day_IP	47158	float64
42	90Day_Oil	47158	float64
43	90Day_GasBoe	47158	float64
44	90Day_ProdDays	47158	float64
45	90Day_IP	47158	float64
46	180Day_Oil	47158	float64
47	180Day_GasBoe	47158	float64
48	180Day_ProdDays	47158	float64
49	180Day_IP	47158	float64
50	270Day_Oil	47158	float64
51	270Day_GasBoe	47158	float64
52	270Day_ProdDays	47158	float64
53	270Day_IP	47158	float64
54	180Day_IP_Corr	44868	float64
55	180Day_Oil_Corr	47158	float64
56	180Day_GasBoe_Corr	47158	float64
57	180Day_ProdDays_Corr	47158	float64

58	GrossProdTime	47158	object
59	ProdDayRatio	47158	object
60	ProdHist	44969	object
61	GrossProdTimeRev	42999	timedelta64[ns]
62	NBRR_Hor_IP_Quintile	5357	category
63	NBRR_Hor_NormBoeCum_Quintile	5500	category
64	CODL_Hor_IP_Quintile	750	category
65	CODL_Hor_NormBoeCum_Quintile	824	category

9.7 List of Jupyter Notebooks

ShapeFiles_Loading_Conditioning	Begin File(s):	WELL.SHP
		DIRECTIONAL_BOTTOMHOLE_LOCATIONS.SHP
		DIRECTIONAL_LINES.SHP
	End File(s):	gdf_surf_aoi (Wells_filtered.shp)
		gdf_dirlines (Dirlines_filtered.shp)
		gdf_bh_aoi (DirBH_filtered.shp)
		surf_bh_mrg_FINAL_CLEAN.pickle
ProductionDataImportMerge	Begin File:	COGCC Production Reports.csv
	Intermediate	Raw File: allproddf
	End File:	allprodaoi_wbh.pickle
ProdDataClean	Begin File:	allprodaoi_wbh.pickle
	End File:	allprodaoi_wbh_dt.pickle
ReassignNCOM	Begin File:	allprod_wbh_dt
	Intermediate Raw File:	allprodaoi_wbh_srtd_Form2.pickle
	End File:	allprodaoi_dt_Final_Clean.pickle
ProdDataCleanPass2	Begin File(s):	allprodaoi_dt_Final_Clean.pickle
	Intermediate Raw File:	allprodaoi_dt_Final_Clean_No_Inj3.pickle
	End File(s):	Prod_DT_Series_Final_WQuantileRank.pickle
GenerateRollup	Begin Files(s):	allprodaoi_dt_Final_Clean_No_Inj3.pickle
	Intermediate Raw File:	
	Prod_DT_Series_Final_WQuantileRank.pickle	
	End_Files(s):	allprodaoi_final_rollup.pickle
RollupExplAnal	Begin Files:	allprodaoi_final_rollup.pickle
TimeSeriesEDA	Begin Files:	Prod_DT_Series_Final_WQuantileRank.pickle
WOE	Begin Files:	allprodaoi_final_rollup.pickle