What kind of cleaning steps did you perform?

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. 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 dataframe (allproddf) consists of 12,922,083 rows and has a file size of over 3.2 GBytes

The shapefile datasets and the area of interest polygon were then loaded using the geopandas module to create geo-dataframes. Each of the three geodataframes were clipped to the area of interest. The procedure for performing this step was to first iterate through the x, y coordinates (geometry) of each well in the geodataframe and if those coordinates fell within the area of interest polygon, it was added to a shortlisted API well list. The shortlisted API well list was then used to filter to only those wells contained in the geodataframe and to create a new dataframe that is filtered to the area of interest. These three new filtered geodataframes 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.

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 and padded with leading zeros instead. 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 would be superceded by corrected information from a later date. This brought the resulting dataset down to 11,910,725 rows.

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 some erroneous Stat_Date values of '3019-12-19' had to be corrected to '2019-12-19 before this conversion was successful.

In addition the prod_days attribute was converted to a timedelta dtype for direct comparison with 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 the many of the wells, prior to an actual formation assignment. To get more accurate production volumes by formation_code, the N-COM code was re-assigned to the actual formation(fm_code). Other wells having a formation_code of N-COM throughout their production history were left intact.

Missing Data

Only a few attributes had missing data. These were the ground_ele, Max TVD, Max MD, MD, and TVD attributes that provide important information about the depth of each well.
How we are filling this in…

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. Examination of the correct TVD values for those wells in the online COGIS database indicated that these values should be clipped at the first 4 digits and the values were corrected accordingly. 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 MD of over 7000 ft and were revised to the MD of the well.

Six outliers were also identified in the Spud_Date attribute, with year values of 2029, 2109 and 2108. These were corrected to 2019 and 2018 respectively.

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

17 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'.

16 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 that had known production through to the end of the production records was 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.

Water volume for one well is extraordinarily large for a producing well. This well is actually a water injection well and the volumes recorded are injection volumes. Therefore all injectors were excluded from the exploratory analysis. Which well is this?

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 (allprodaoi). This reduced the total number of rows in the dataframe to 6,442,494 increasing the columns to 72 The resulting dataframe name is allprodaoidt.

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, producing a total of 86 columns.

Removing Unwanted Columns and Renaming Others

Thirteen 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 frame. In addition, two additional feature attributes were calculated that were pertinent to the time-series data. These features will be discussed in the feature engineering section of this report.

Final Production TimeSeries DataFrame

The final production time-series dataframe has a total of 6,442,494 rows and 77 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
1999-08-01	1999-08-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-10-01	1999-10-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-06-01	1999-06-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-03-01	1999-03-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-07-01	1999-07-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-05-01	1999-05-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-02-01	1999-02-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW
1999-01-01	1999-01-01	1231434000	12314340	00	Vertical	47120	KERR MCGEE OIL & GAS ONSHORE LP	41385	HS RESOURCES INC ...	5-28A HSR-RENSHAW

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 has 37,299 rows (the number of unique wells in the final production time-series dataframe) and 49 columns and is 13. 7 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	well_type_cat	well_type_cat2	OpCurNum1	OpCurName1	OpHistNum1	OpHistName1	OpCurNum2	OpCurName2	OpHistNum2	OpHistName2
0	10502900	Vertical	Non-Horizontal	72085	PETRO-CANADA RESOURCES (USA) INC	94090	WALSH PRODUCTION INC ...	72085	PETRO-CANADA RESOURCES (USA) INC	94090	WALSH PRODUCTION INC ...
1	10504400	Vertical	Non-Horizontal	72085	PETRO-CANADA RESOURCES (USA) INC	94090	WALSH PRODUCTION INC ...	72085	PETRO-CANADA RESOURCES (USA) INC	94090	WALSH PRODUCTION INC ...
2	10507000	Vertical	Non-Horizontal	95620	WESTERN OPERATING COMPANY	96155	WHITING PETROLEUM CORP ...	95620	WESTERN OPERATING COMPANY	95620	WESTERN OPERATING COMPANY
3	10524200	Vertical	Non-Horizontal	10330	INVESTMENT EQUIPMENT LLC	39150	HEARTLAND OIL & GAS COMPANY ...	10330	INVESTMENT EQUIPMENT LLC	10330	INVESTMENT EQUIPMENT LLC
4	10526300	Vertical	Non-Horizontal	59100	MONAHAN* REX FAMILY TRUST	59100	MONAHAN* REX ...	59100	MONAHAN* REX FAMILY TRUST	59100	MONAHAN* REX ...
5	10528900	Vertical	Non-Horizontal	10330	INVESTMENT EQUIPMENT LLC	39150	HEARTLAND OIL & GAS COMPANY ...	10330	INVESTMENT EQUIPMENT LLC	10330	INVESTMENT EQUIPMENT LLC
6	10529900	Vertical	Non-Horizontal	10330	INVESTMENT EQUIPMENT LLC	39150	HEARTLAND OIL & GAS COMPANY ...	10330	INVESTMENT EQUIPMENT LLC	10330	INVESTMENT EQUIPMENT LLC
7	10532100	Vertical	Non-Horizontal	46290	KP KAUFFMAN COMPANY INC	46290	K P KAUFFMAN COMPANY INC	46290	KP KAUFFMAN COMPANY INC	46290	K P KAUFFMAN COMPANY INC

Fig 4: Slice of final rollup-dataframe

-Data and code need to be uploaded to Github