Greater Panhandle Wind Generation Data Analysis

Austin Shinn

Packages Used:

```
library(tidyverse)
library(tidymodels)
library(timetk)
library(skimr)
library(lubridate)
library(janitor)
library(modeltime)
library(cowplot)
library(plotly)
```

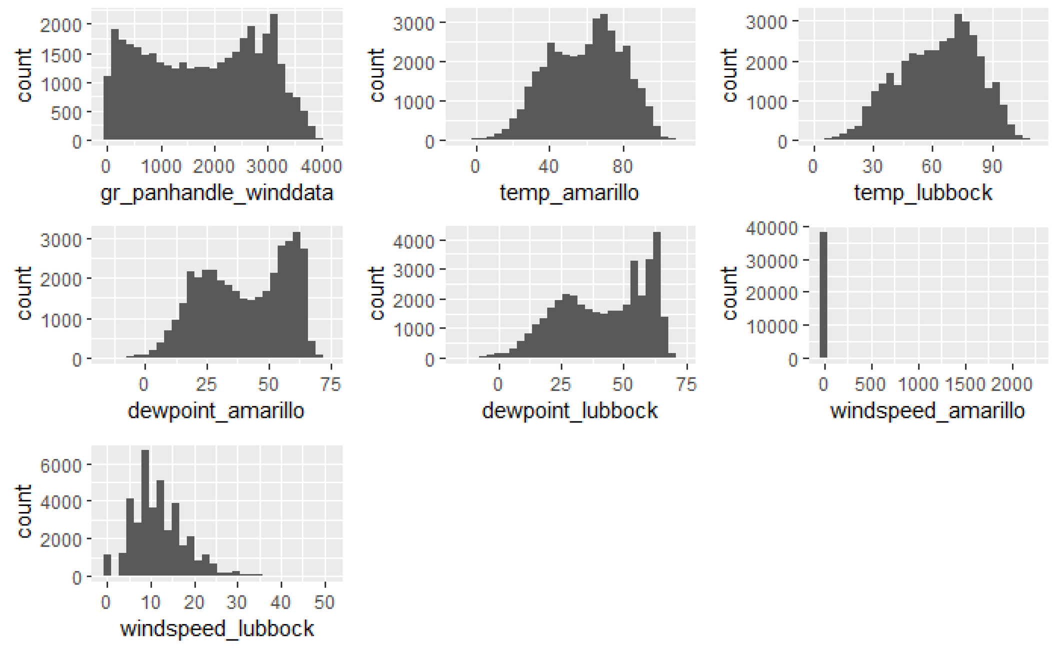
Pulling and cleaning data from Excel sheets:

```
windcapacity_data <- read_csv("data/panhandlewindcapacity1.csv") %>%
  clean_names() %>%
  rename(amarillo_or_lubbock = closer_to_a_or_l)

wind_data <- read_csv("data/windsampledata.csv") %>%
  clean_names() %>%
  select(-market_day, -year) %>%
  mutate(
         datetime = mdy_hm(datetime),
         month = factor(month, levels = c("JANUARY", "FEBRUARY", "MARCH", "APRIL", "MAY", "JUNE", "JULY", "AUGUS
T", "SEPTEMBER", "OCTOBER", "NOVEMBER", "DECEMBER"))
         ) %>%
  distinct(datetime, .keep_all = TRUE)

#wind_data <- wind_data_prelim[!(wind_data_prelim$year == 2017),] %>%
#  select(-year)
```

Possibility: taking out 2017 data since the exact dates of construction of plants weren't available. ran it this way (taking out all 2017) and the produced models were less accurate. More data seems to outweigh the cons of slightly inaccurately adjusted data in regards to total possible production, so I ended up keeping all data from 2017 for model training.

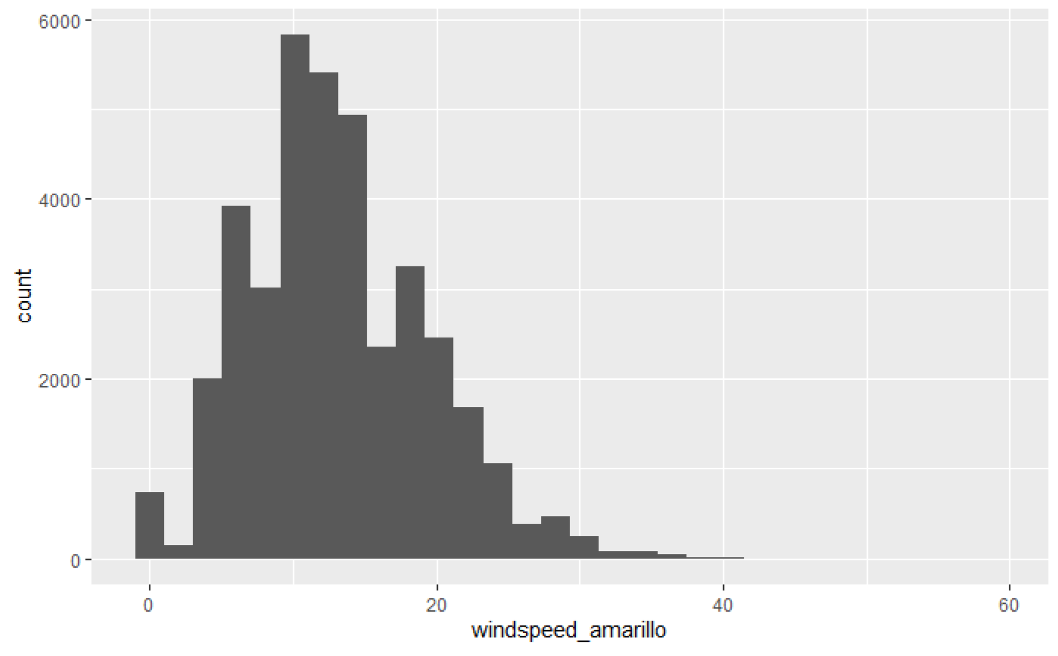
```
-- Data Summary ------------------------
                           Values
Name                       wind_data
Number of rows             38444
Number of columns          13
_______________________
Column type frequency:
  character                1
  factor                   1
  numeric                  10
  POSIXct                  1
________________________
Group variables            None

-- Variable type: character ----------------------------------------------------------------------------------------
-----------------------------------
# A tibble: 1 x 8
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
* <chr>             <int>         <dbl> <int> <int> <int>    <int>      <int>
1 peak_type             0             1     6     7     0        3          0

-- Variable type: factor -------------------------------------------------------------------------------------------
-----------------------------------
# A tibble: 1 x 6
  skim_variable n_missing complete_rate ordered n_unique top_counts
* <chr>             <int>         <dbl> <lgl>      <int> <chr>
1 month                 0             1 FALSE         12 JUL: 3720, AUG: 3720, JUN: 3600, SEP: 3600

-- Variable type: numeric ------------------------------------------------------------------------------------------
-----------------------------------
# A tibble: 10 x 11
   skim_variable             n_missing complete_rate      mean       sd     p0    p25     p50     p75    p100 hi
st
 * <chr>                         <int>         <dbl>     <dbl>    <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl> <c
hr>
 1 temp_amarillo                   155         0.996    59.4      20.0    -9.9   44.1    62      74     109     ▁
▃▇▇▂
 2 temp_lubbock                     60         0.998    62.6      19.4     0     48      64.9    77     111     ▁
▃▇▇▂
 3 dewpoint_amarillo               182         0.995    40.2      17.4   -16.1   26      41      55.9    73     ▁
▃▇▇▇
 4 dewpoint_lubbock                 60         0.998    42.0      17.5   -16.1   27      44.1    57.9    71.6   ▁
▂▇▆▇
 5 windspeed_amarillo              197         0.995    14.3      51.2     0      9      12.7    17.3  2237.    ▇
▁▁▁▁
 6 windspeed_lubbock                64         0.998    11.7       5.99    0      8      10.4    15      50.6   ▇
▇▂▁▁
 7 hour_ending                       0         1        12.5       6.92    1      7      13      19      24     ▇
▇▇▇▇
 8 precipitation_lwc_lubbock     36042         0.0625    0.0686    0.249  0.01   0.01    0.01    0.02    5.54  ▇
▁▁▁▁
 9 precipitation_lwc_amarillo    36496         0.0507    0.0966    0.307  0.01   0.01    0.02    0.05    4.91  ▇
▁▁▁▁
10 gr_panhandle_winddata             3         1.00   1777.     1083.     -0.28 784.   1813.   2726.   4104.    ▇
▇▇▇▂

-- Variable type: POSIXct ------------------------------------------------------------------------------------------
-----------------------------------
# A tibble: 1 x 7
  skim_variable n_missing complete_rate min                 max                 median              n_unique
* <chr>             <int>         <dbl> <dttm>              <dttm>              <dttm>                 <int>
1 datetime              0             1 2017-06-01 00:00:00 2021-10-19 23:00:00 2019-08-10 23:30:00    38444
```

There is missing data for most predictors, but not to a great enough extent that it would be a problem if they were simply removed. the exception is precipitation for both Amarillo and Lubbock. Models will be run both with and without the few non-NA values to see if it's worth keeping/imputing.



The default scale of the histogram of wind speed in Amarillo looks way off. We can see that there is an outlier value in the mid-2000s for wind speed, which has to be an error in measurement or recording.

There were 18 instances where the wind speed readings measured 2236.716, so we arrange wind_data by Amarillo wind speed in descending order then remove the first 18 readings, since we're removing columns with NA values anyway. With the misinput data removed, the Amarillo wind speed plot looks like this.



Data is split into training and testing sets with an 80/20 split, which is suitable for larger datasets. k-fold cross validation with 5 folds and 3 repeats is used to keep the data from becoming too biased to the training set.

Preprocessing steps:

Near-zero variance filter removes variables that are sparse and unbalanced, meaning variables that may have basically the same value for all observations. I don't think this was necessary because the data is so varied, but I just kept it because it doesn't hurt.

Yeo-Johnson transformation reduces skew of variables, which I used on all predictors for temperature, dewpoint temp, and wind speed. It's helpful for some, but not necessary for other types of models.

Removed datetime variable because this is not a time-series forecasting machine learning model.

All nominal / factor variables are changed to dummy variables (binary) which is better for many models.

datetime <S3: POSIXct>	temp_amarillo <dbl>	temp_lubbock <dbl>	dewpoint_amarillo <dbl>	dewpoint_lubbock <dbl>
2019-03-13 14:00:00	86.49835	129.59321	29.038726	42.336617
2019-03-13 12:00:00	80.52240	121.61752	33.244607	45.016751
2019-03-13 15:00:00	85.15051	132.44570	27.988538	39.668970
2019-03-13 16:00:00	87.99811	132.44570	29.984349	37.014505
2019-03-13 17:00:00	83.50568	129.59321	29.038726	34.374003
2020-10-28 23:00:00	44.16523	70.58864	33.244607	42.336617
2020-10-28 21:00:00	42.89497	66.22170	33.244607	41.134591
2019-03-13 18:00:00	71.77914	127.26801	33.244607	35.560447
2020-10-29 00:00:00	42.89497	72.67498	33.244607	41.134591
2020-07-23 13:00:00	109.37075	208.28123	65.089860	83.584762

1-10 of 30,533 rows | 1-5 of 24 columns Previous 1 2 3 4 5 6 ... 100 Next

Models tested and results (for both optimal hyperparameters and actual model performance):

k-nearest neighbors best model: neighbors = 11, RMSE = 601.17

random forest best model: mtry = 6, min_n = 2, RMSE = 546.40

boosted tree: mtry = 10, min_n = 11, learn_rate = .631, RMSE = 587.55

single-layer neural network: hidden units = 5, penalty = 1.00, RMSE = 836.2845

mars model: number of terms = 81, prod_degree = 2, RMSE = 606.0637

The best performing model by RMSE (and also R^2) was the random forest model. The following is a graph of predicted values from the model and actual values from the dataset (use the slider to zoom into a specific timeframe). Overall, I would say the model does a good job of predicting the trends of the actual data, and part of the model error was in events that could not be predicted. At least a couple times, actual value falls to near-zero or actually zero when the model predicts a higher number, which I would assume is equipment failure or maintainence. The variables used in the model are also ones readily available from public weather forecasting data, which makes it realistic in practical usage for wind generation forecasting. Inclusion of additional relevant variables may further increase model accuracy.