energy_factor_analysis

December 12, 2021

[25]:
```
import pandas as pd
import os
import statsmodels.api as sm
from sklearn import linear_model
import numpy as np
```

0.1 This notebooks explores the relationship between a state's number of vehicle registrations, population, GDP per capita, GDP per capita by industry, C02 emissions, average yearly tempature, average yearly windspeed, minimum yearly tempature, maximim yearly tempature, total yearly precipitation , and total yearly snowfall on it's energy consumption.

0.1.1 The goal is to model a US state's energy consuption by using the data listed above. With this model we can make energy consuption predictions and understand what leads to high energy consuption.

0.1.2 The contents of the notebook include

- **Data Gathering**
  - read in the dataframes that have been cleaned by data_gathering_and_cleaning notebook
- **Data analysis**
  - create a multiple linear regression model for energy consuption
- **Conclusion**
  - Discuss what we discovered and draw conclusions

Note: If there are no files in the Data/cleaned diretory, you will need to run the 'data_gathering_and_cleaning" notebook to clwan and write out the files to that directory.

0.1.3 Data Gathering

This section of the notebooks reads in the data files and stores them im pandas dataframes. The dataframes frames in this section all have columns of represting years ranging from [1967-2020] and rows for each state.

[34]:
```
csv_path = os.path.join(os.getcwd(), "data/cleaned/csv")
excel_path = os.path.join(os.getcwd(), "data/cleaned/excel")
```

[44]:
```
#Read in all datasets here
vehicle_registration_df = pd.read_csv(os.path.join(csv_path,␣
 ↪"vehicle_registrations_by_state.csv"))
energy_consumption_per_real_gdp_df = pd.read_csv(os.path.join(csv_path,␣
 ↪"energy_consumption_per_real_gdp.csv"))
current_dollar_gdp_df = pd.read_csv(os.path.join(csv_path, "Current_dollar_GDP.
 ↪csv")) #in millions
total_consuption_df = pd.read_csv(os.path.join(csv_path, "total_consuption.
 ↪csv")) #in million Btu
industy_gdp_by_state_df = pd.read_csv(os.path.join(csv_path,␣
 ↪"industy_gdp_by_state.csv"))
total_population_df = pd.read_csv(os.path.join(csv_path, "total_population.
 ↪csv"))
real_gdp_df = pd.read_csv(os.path.join(csv_path, "real_GDP.csv")) #in millions
co2_emissions_df = pd.read_excel(os.path.join(excel_path, "co2_emissions.xlsx"))
tavg_df =  pd.read_csv(os.path.join(csv_path + '/NOA', "TAVG.csv"))
wind_df =  pd.read_csv(os.path.join(csv_path + '/NOA', "AWND.csv"))
tmax_df =  pd.read_csv(os.path.join(csv_path + '/NOA', "TMAX.csv"))
tmin_df =  pd.read_csv(os.path.join(csv_path + '/NOA', "TMIN.csv"))
precip_df = pd.read_csv(os.path.join(csv_path + '/NOA', "PRCP.csv"))
snow_df = pd.read_csv(os.path.join(csv_path + '/NOA', "SNOW.csv"))
```

[45]:
```
#Use the columns that are in each dataframe after columns with empty values␣
 ↪have been dropped.
columns_to_evaluate = list(set(vehicle_registration_df.columns).
 ↪intersection(total_population_df.columns).intersection(total_consuption_df.
 ↪columns).intersection(real_gdp_df.columns).
 ↪intersection(industy_gdp_by_state_df.columns).intersection(co2_emissions_df.
 ↪columns).intersection(tavg_df.columns).intersection(wind_df.columns).
 ↪intersection(tmax_df.columns).intersection(tmin_df.columns).
 ↪intersection(precip_df.columns).intersection(snow_df.columns))
columns_to_evaluate
```

[45]:
```
['2013',
 '2017',
 '2012',
 '2011',
 '2015',
 '2016',
 '2009',
```

```
 '2008',
 '2007',
 '2014',
 '2018',
 '2010',
 'Unnamed: 0']
```

[46]:
```python
#ensure each column we are going to evaluate has the same number of values
for col in columns_to_evaluate:
    if(not (len(vehicle_registration_df[col]) == len(total_consuption_df[col])␣
 ↪== len(total_population_df[col]) == len(real_gdp_df[col])==␣
 ↪len(industy_gdp_by_state_df[col]) == len(co2_emissions_df[col]) ==␣
 ↪len(tavg_df[col])== len(wind_df[col])== len(tmax_df[col])==␣
 ↪len(tmin_df[col])== len(precip_df[col])== len(snow_df[col]))):
        print("unequal entries for column:" + col)
```

0.1.4 Data Analysis

This section of the notebooks creates a multiple linear regression model for a state's energy consuption.

In the model summary each variable is represented by the following

- x1: Vehicle regisrations
- x2: Population
- x3: GDP per capita
- x4: Industry GDP per capita
- x5: C02 emissions
- x6: Average tempature
- x7: Average wind speed
- x8: Maximum tempature
- x9: Minimum tempature
- x10: Total precipitation
- x11: Total snow fall

There are some other values in the summary that give us a good indication as to how well our model fits energy consuption such at the r squared value and F statistic.

[47]:
```python
# loop through the data frames and add each value to data_point_pairs array.
# The data_point_pairs array will be the
# [vehicle registration, population, GDP, Industry GDP, CO2 emissions, average␣
 ↪tempature, average wind speed, max temperature, min tempature, total␣
 ↪precipitation, total snowfall]
# value for each year and each state
# The total_consumption_vals will be the cooresponding energy consuption value
# for the data point pairs item
data_point_pairs = []
total_consumption_vals = []
for col in columns_to_evaluate:
```

```
    for i in range(0,50):
        pair = [vehicle_registration_df.iloc[i][col], total_population_df.
↪iloc[i][col], real_gdp_df.iloc[i][col], industy_gdp_by_state_df.
↪iloc[i][col], co2_emissions_df.iloc[i][col],tavg_df.iloc[i][col],wind_df.
↪iloc[i][col],tmax_df.iloc[i][col],tmin_df.iloc[i][col],precip_df.
↪iloc[i][col],snow_df.iloc[i][col]]
        data_point_pairs.append(pair)

        total_consumption_vals.append(total_consuption_df.iloc[i][col])
```

[48]:
```
print("vehicle registration:" , data_point_pairs[0][0])
print("population: ", data_point_pairs[0][1])
print("GDP: ", data_point_pairs[0][2])
print("Industry GDP: ", data_point_pairs[0][3])
print("C02 emissions: ", data_point_pairs[0][4])
print("Average tempature:" ,data_point_pairs[0][5])
print("Average Wind Speed:" ,data_point_pairs[0][6])
print("Maximim tempature:" ,data_point_pairs[0][7])
print("Mimimum tempature:" ,data_point_pairs[0][8])
print("Total Precipitation:" ,data_point_pairs[0][9])
print("Total snowfall:" ,data_point_pairs[0][10])
print("total energy consuption:" ,total_consumption_vals[0])
```

```
vehicle registration: 4787219.0
population:  738.0
GDP:  54748.0
Industry GDP:  11241.679347826086
C02 emissions:  121.1630059889289
Average tempature: 6.774734157214605
Average Wind Speed: 3.8361776859504135
Maximim tempature: 14.073909594750376
Mimimum tempature: -0.4421591745467444
Total Precipitation: 46.89005639838973
Total snowfall: 190.3120019711779
total energy consuption: 597975.0
```

[49]:
```
X = data_point_pairs
y = total_consumption_vals
lm = linear_model.LinearRegression()
model = lm.fit(X,y)

#predict energy consuption for vehicle registration = 4610845 , population =699␣
↪(10,000), GDP = 55911,
#Industry GDP = 9717, C02 emissions = 121, Average tempature = 6.7, Average␣
↪Wind Speed = 2.5
#Maximim tempature = 14.07, Mimimum tempature = -0.44, Total Precipitation =␣
↪47, Total snowfall: 190
```

```python
predictions = lm.predict([[4610845, 699, 55911, 9717, 121, 6.7, 2.5, 14.07, -0.
  ↪44, 47, 190]])
print("Predicted energy consumpion:", predictions )

model = sm.OLS(y, X).fit()
model.summary()
```

```
Predicted energy consumpion: [1181779.00795566]
```

[49]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      y   R-squared (uncentered):
0.967
Model:                            OLS   Adj. R-squared (uncentered):
0.967
Method:                 Least Squares   F-statistic:
1716.
Date:                Mon, 06 Dec 2021   Prob (F-statistic):
0.00
Time:                        17:09:05   Log-Likelihood:
-9442.2
No. Observations:                 650   AIC:
1.891e+04
Df Residuals:                     639   BIC:
1.896e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0918      0.011     -8.176      0.000      -0.114      -0.070
x2           172.2853     18.957      9.088      0.000     135.059     209.512
x3             0.6571      0.307      2.140      0.033       0.054       1.260
x4           -17.2976      2.169     -7.976      0.000     -21.556     -13.039
x5          1.388e+04    357.388     38.825      0.000    1.32e+04    1.46e+04
x6           -1.3e+05   4.22e+05     -0.308      0.758   -9.59e+05    6.99e+05
x7          -3.16e+04   6312.268     -5.006      0.000    -4.4e+04   -1.92e+04
x8          6.193e+04   2.11e+05      0.293      0.770   -3.53e+05    4.77e+05
x9          8.208e+04   2.11e+05      0.390      0.697   -3.31e+05    4.96e+05
x10          -32.0556    749.195     -0.043      0.966   -1503.236    1439.125
x11          234.6605    280.403      0.837      0.403    -315.962     785.283
==============================================================================
Omnibus:                       39.487   Durbin-Watson:                   2.276
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              131.060
```

```
Skew:                           0.156   Prob(JB):                     3.47e-29
Kurtosis:                       5.178   Cond. No.                     1.90e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.9e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

0.1.5 Conclusion

This section of the notebooks discusses the results Looking at the United states as a whole, the most prominent effects on enery consumption are C02 emissions, average tempature, average wind speed, maximum tempature, minimum tempature and total precipitation.

Avergage tempature and wind speed both have a large negative effect on energy consuption. For every increase in degrees celcuis, the expected energy consuption is expected to drop by 130,000 million british thermal units (BTU). The energy needed for heating during cold tempatures could accout for this relationship. Similarly, for every increase in 1 mile per hour wind speed, the expected energy consumption is expected to drop by 316,000 BTU. This seems counter intutaive since more wind speed would increase the amount of wind energy consumed. However, its probable that increasing more win energy significantly decreases the amount of other energy sources that are consumed.

Both maximum and mimimum yearly tempature have a large positive impact on energy consuption, with 61,930 and 82,080 incease in million BTU per increase in degrees celcius, respectively. This result is contradictory to the predictions that the average tempature has an inverse relationship with energy consuption, so it is possible we are missing something to explain this inconsistency.

C02 emissions also have a large positive effect on energy consumption. For every increase in 1 million metric ton, energy consuption is expected to increase by 13,880 BTU. This is expected since the more energy you consume, the more C02 emissions you produce.