Predoc Coding Task 2025

1 Question 1

There are 1121 distinct values of the resource name variable and 194 distinct values of the QSE variable.

2 Question 2

In ERCOT, a QSE (Qualified Scheduling Entity) is a company that helps power plants and electricity buyers take part in the Texas electricity market. It handles tasks like sending schedules to ERCOT, managing energy sales, and making sure power gets delivered when promised.

3 Question 3

There are 1127 distinct QSE/resource name pairs.

3.1 (a)

Yes, many QSEs manage multiple unique resource names. This indicates that QSEs act as central coordinators for several power plants or generators. A single QSE typically represents many resources in the ERCOT market, handling their energy scheduling and dispatch.
The top 10 highest frequency QSEs are:

Table 1: QSEs with highest resource name frequencies

QSE	Resource name count
QTENSK	173
QLUMN	75
QNRGTX	50
QCALP	44
QECNR	40
QAEN	34
QLCRA	32
QCPSE	25
QTEN23	24
QSHEL2	22

3.2 (b)

Yes, a single resource name can be paired to more than one QSE over time. 6 resource names are connected to more than one QSE. The overarching pattern over time is that all these 6 resource names switched from QENEL5 QSE to another QENEL QSE soon after a day/few days, possibly because of change in ownership or better results with the new QSE.

4 Question 4

4.1 (a)

The resource type variable takes 15 unique, non missing values.
definitions are probably as follows:

- DSL – Diesel-fired generators
- WIND – Wind turbines, a major renewable resource in ERCOT
- HYDRO – Conventional hydroelectric generation

- NUC – Nuclear power generation, e.g., South Texas Project
- RENEW – Catch-all for renewable sources not individually categorized (e.g., biomass)

4.2 (b)

There are 4 empty strings in the resource type column. The resource names missing their resource types are:

- GALLOWAY_SOLAR1
- ROSELAND_SOLAR3
- SSPURTWO_WIND_1
- SWEETWN2_WND24

I filled the first 2 empty strings with PVGR for photovoltaic solar energy generation and the last 2 with WIND for wind turbine energy generation.

5 Question 5

The merged file was produced as instructed and has been uploaded.

6 Question 6

In plot (a) given by Figure 1, where output is summed by day, we observe clear variability in total generation across different days. Some days show significant peaks in output, notably around early February, which might reflect spikes in electricity demand or increased supply availability. There are also noticeable dips, especially in the days following early February, which could be due to factors such as reduced demand (possibly on weekends), unfavorable weather conditions affecting renewable sources, or scheduled maintenance on generating units.

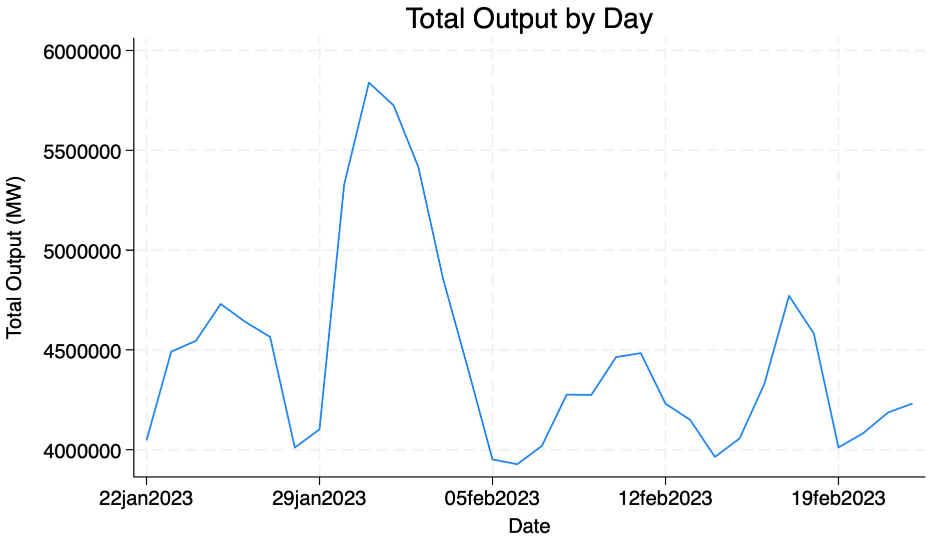

Figure 1: Total Output by Day

Plot (b) given by Figure 2, which presents output summed by hour of day (0–23), reveals a classic daily load curve. Output is at its lowest during the early morning hours between midnight and 4 AM, gradually rising as the morning progresses. A steep increase begins around 5 or 6 AM, with a sustained peak between 8 AM and noon, corresponding to the start of typical working hours. There's another peak in the evening hours from around 5 PM to 8 PM, likely reflecting residential electricity use as people return home. This pattern then declines late at night, showing how demand shapes generation patterns throughout the day.

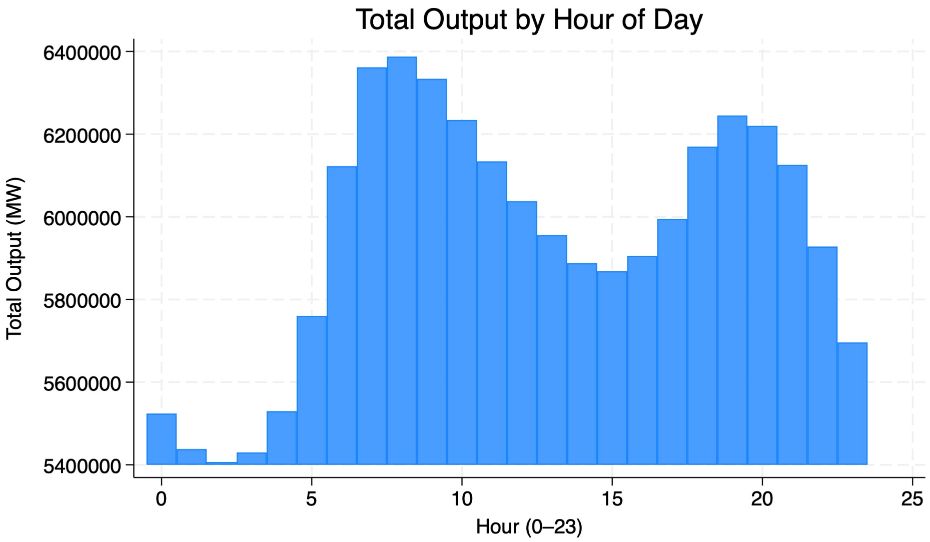

Figure 2: Total Output by Hour of Day

In plot (c) given by Figure 3, where output is summed by both hour of day and fuel type, the roles of different fuel sources in meeting demand become clearer. Coal and nuclear exhibit relatively constant output throughout the day, indicative of their use as baseload generators. In contrast, natural gas shows marked increases during morning and evening peaks, suggesting its role as a flexible, load-following source that adjusts to short-term demand fluctuations. Solar generation rises sharply during mid-day hours, peaking between 10 AM and 2 PM, and falling to zero during nighttime—reflecting the availability of sunlight. Wind output tends to increase during the late evening and early morning hours, consistent with typical wind patterns. Fuel types categorized as "Other" contribute marginally. Together, these plots reflect a grid that relies on a balanced mix of stable and variable generation sources to meet hourly and daily electricity needs.

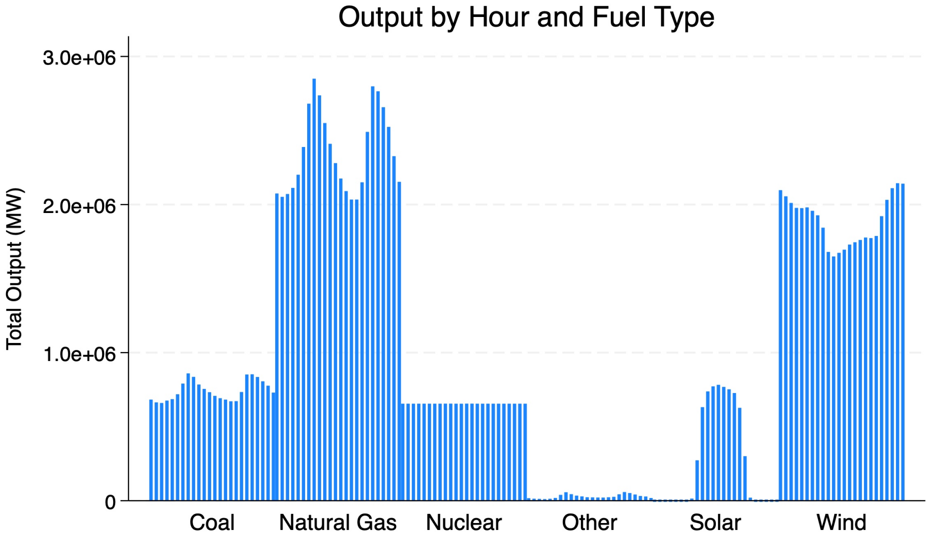

Figure 3: Output by Hour and Fuel Type

7 Question 7

No, the data in 6(a) does not appear stationary and appears to have some periodicity, as the output fluctuates from day to day.

Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
-2.138	-3.709	-2.983	-2.623

Table 2: ADF Unit Root Test for `telemetered_net_output` (lags = 0, no drift). *MacKinnon approximate p-value for Z(t): 0.2296*

In this case as shown in Table 2, the test statistic is -2.138, which is higher (less negative) than the 5% critical value of -2.983. Additionally, the MacKinnon approximate p-value is 0.2296, which is well above the conventional significance threshold of 0.05. Together, these indicate that we fail to reject the null hypothesis of a unit root. In simpler terms, this suggests that the output series behaves like a random walk and is non-stationary, meaning its statistical properties—such as mean and variance—change over time.
Now we do first differencing on the data and again perform the unit root test for non-stationary data.

Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
-3.635	-3.716	-2.986	-2.624

Table 3: ADF Unit Root Test for `d_output` (lags = 0, no drift). *MacKinnon approximate p-value for Z(t): 0.0051*

The test statistic is -3.635 as shown in Table 3, which is more negative than the 5% critical value of -2.986, and the p-value is 0.0051—well below the conventional threshold of 0.05. This allows us to reject the null hypothesis of a unit root and conclude that the differenced series is stationary. In other words, while the original output data showed signs of non-stationarity, its first difference does not, suggesting that the data is integrated of order one, I(1).

8 Question 8

Variable	Coefficient	Std. Err.	z	P¿|z|	[95% Conf. Interval]
Constant	184824.000	5521.044	33.48	0.000	[174003.000, 195645.100]
AR(1)	2.215	0.0246	89.96	0.000	[2.1668, 2.2634]
AR(2)	-1.747	0.0479	-36.44	0.000	[-1.8411, -1.6531]
AR(3)	0.512	0.0278	18.41	0.000	[0.4579, 0.5670]
Sigma	2565.034	58.749	43.66	0.000	[2449.888, 2680.181]

Table 4: ARIMA Regression Results for `telemetered_net_output` (Sample: 1 to 768)

The AR(3) model given in Table 4for hourly electricity output indicates a strong autoregressive structure in the data. All three lag terms are statistically significant at the 1% level. The first lag has a positive and large coefficient (2.215), suggesting that current output is heavily influenced by the immediately preceding hour. The second lag has a large negative coefficient (1.747), and the third lag is again positive (0.512). This alternating sign pattern suggests cyclical behavior in electricity output. The overall model fit is strong, as indicated by the highly significant Wald chi-squared statistic (p ¡ 0.001). However, the very high coefficient for the first lag and the fact that the sum of the AR coefficients is close to 1 may suggest persistent autocorrelation, meaning shocks to the system have a lasting effect. If residuals still show significant autocorrelation, the AR(3) model may be underfitting, and incorporating additional lags, exogenous variables (like fuel type, hour-of-day effects), or moving average components (ARMA/ARIMA) might improve the model. But as it stands, the AR(3) model captures key temporal dependencies in the hourly output data.

9 Question 9

9.1 (a)

Table 5: Regression of Telemetered Output on Fuel Type

Variable	Coefficient	Std. Err.	t	[95% Conf. Interval]
Natural Gas	-161.332***	0.401	-402.57	[-162.118, -160.547]
Nuclear	416.110***	0.792	525.32	[414.557, 417.662]
Other	-221.911***	0.419	-529.90	[-222.732, -221.090]
Solar	-209.237***	0.417	-501.63	[-210.054, -208.419]
Wind	-183.023***	0.398	-460.21	[-183.802, -182.243]
Constant	223.542***	0.384	581.76	[222.789, 224.295]

Notes: ***p < 0.01, **p < 0.05, *p < 0.1.
Number of observations = 3,008,438
F(5, 3008432) > 99999.00
Prob > F = 0.0000
R-squared = 0.2641
Adjusted R-squared = 0.2641
Root MSE = 108.59

The regression output in Table 5 shows how electricity output (as measured by telemetered_net_output) varies across different fuel types, with the coal category omitted. The intercept, or constant term, represents the mean output for the coal fuel category. The coefficients for each fuel type indicate how their average output differs from this base category. For instance, units using Nuclear fuel produce, on average, 416.11 MW more than the base category, which is a large and statistically significant positive difference. Conversely, units powered by Natural Gas, Other fuels, Solar, and Wind all show significantly lower average outputs than the base: by 161.33 MW, 221.91 MW, 209.24 MW, and 183.02 MW respectively. All these differences are highly statistically significant (p ¡ 0.001), suggesting that the type of fuel used is strongly associated with average electricity output. These results could reflect both the inherent generation capacities of each technology and their roles in the energy mix — for example, nuclear tends to operate at high and steady output levels, while renewables like solar and wind are more variable and often have lower capacity factors.

9.2 (b)

Table 6: Regression of Telemetered Output on Day of the Week

Variable	Coefficient	Std. Err.	t	[95% Conf. Interval]
Monday	3.490***	0.261	13.36	[2.978, 4.002]
Tuesday	4.713***	0.261	18.04	[4.201, 5.225]
Wednesday	5.554***	0.261	21.28	[5.042, 6.066]
Thursday	6.239***	0.277	22.53	[5.697, 6.782]
Friday	6.231***	0.277	22.50	[5.688, 6.773]
Saturday	3.124***	0.277	11.28	[2.581, 3.666]
Constant	43.332***	0.185	234.57	[42.970, 43.694]

Notes: ***p < 0.01, **p < 0.05, *p < 0.1.
Number of observations = 3,008,438
F(6, 3,008,431) = 135.55
Prob > F = 0.0000
R-squared = 0.0003
Adjusted R-squared = 0.0003
Root MSE = 126.57

The regression results in Table 6 examine how electricity output varies across the days of the week, using Sunday as the baseline category. The constant term of approximately 43.33 represents the average electricity output on Sundays. Compared to this, output is higher on all other days of the week. On Mondays, output increases by about 3.49 MW, and continues to rise through the week, peaking on Thursdays and Fridays, when the output is about 6.24 MW higher than on Sundays. Saturdays see a smaller increase of 3.12 MW, indicating a slight weekend dip relative to weekdays. All coefficients are statistically significant at the 1% level, as indicated by the very low p-values. However, despite this significance, the overall explanatory power of the model is minimal—reflected in the R-squared value of 0.0003—suggesting that while day-of-week has some predictive value, most of the variation in electricity output is driven by other factors. The pattern observed is consistent with typical work-week electricity demand cycles, with output rising on weekdays due to increased industrial and commercial activity and dipping slightly on weekends.

9.3 (c)

Table 7: Regression of Telemetered Output on Week

Variable	**Coefficient**	**Std. Err.**	**t**	**[95% Conf. Interval]**
Week 5	6.658***	0.221	30.09	[6.225, 7.092]
Week 6	-3.197***	0.221	-14.47	[-3.630, -2.764]
Week 7	-2.157***	0.221	-9.77	[-2.590, -1.724]
Week 8	-3.954***	0.259	-15.28	[-4.461, -3.447]
Constant	47.652***	0.157	303.91	[47.344, 47.959]

Notes: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.
Number of observations = 3,008,438
F(4, 3,008,433) = 698.41
Prob > F = 0.0000
R-squared = 0.0009
Adjusted R-squared = 0.0009
Root MSE = 126.53

The regression output in Table 7 shows the results of regressing electricity output on dummy variables for weeks 5 to 8, with an omitted reference week 4. Each coefficient indicates the average difference in output for that week compared to the base week. The constant term (_cons) of 47.65 represents the average output in the base week. Week 5 had a significantly higher output, averaging 6.66 units more than the base week. Conversely, weeks 6, 7, and 8 show lower output, with decreases of 3.20, 2.16, and 3.95 units, respectively, compared to the base week. All coefficients are statistically significant at the 1% level, suggesting these differences are unlikely due to chance. However, the R-squared is only 0.0009, indicating that week-to-week variation explains less than 0.1% of the total variation in output. This suggests that while some weeks see higher or lower output, week is not a strong predictor, and there are likely many other factors (like fuel type, time of day, weather, etc.) driving generation levels.The weekly variation in electricity output reflected by the regression coefficients can be attributed to several underlying factors. Changes in weather conditions across weeks such as variations in temperature or solar radiation can significantly influence the output from renewable sources like solar and wind. Additionally, fluctuations in electricity demand, driven by industrial cycles, or residential consumption patterns can lead to changes in how much electricity is generated in a given week.