1.

Window size	steps	MSE
12	15	3.6602
20	15	2.824
20	5	1.5627

When the window size is larger than steps, the MSE decreases. When the steps become smaller, it means that the overlap rate becomes higher and the training effect is relatively better, but overfitting may also occur.

2.

(i) After adding Volume, the effect of the model is greatly reduced, and the MSE rises directly from 1.5627 to 261.3687. This may be because the differences between volumes are too great and the model is not complex enough to learn such complex features.

(ii)

Open	High	Low	Close	Volume	MSE
1	1	1	1	1	261.3687
1	1	1	1	0	1.5627
0	1	1	1	1	357.9577
1	1	1	0	0	1.5354
0	1	1	1	0	1.4987
0	1	1	0	0	2.2364
0	1	0	1	0	4.0436
1	1	0	0	0	1.5361
1	1	0	1	0	1.9700
0	1	1	0	1	238.8241

3.

Open	High	Low	Close	Volume	Normalize	MSE
0	1	1	0	1	0	238.8241
0	1	1	0	1	1	3.6681
0	1	1	1	0	0	1.4987
0	1	1	1	0	1	2.6684
1	1	1	1	0	0	1.5627
1	1	1	1	0	1	1.9216
1	1	1	1	1	0	261.3687
1	1	1	1	1	1	1.8684

For features as different as Volume, normalizing the model will be of great help. Normalize differences. However, the effect is not so obvious when it comes to data that is not very different or that is originally similar to a normal distribution. It can be clearly seen from the above table that when the feature has a Volume, the MSE drops a lot. But it doesn't have much impact on other features.

4.
The window size should be greater than or equal to the step size, not less. The window size defines the number of consecutive data points in each input sample, while the step size defines the jump in data points from one sample to the next. If the window size is smaller than the step size, it will cause gaps between data samples, and important information may be missed. Typically, the step size is smaller than or equal to the window size to ensure data continuity and coverage. (I have done an experiment on the first question)

5.
Time shifting: Shifting the time series signal by a fixed time interval on the time axis can increase the diversity of the data set.
Noise addition: Adding random noise to time series signals can increase the robustness of the data set.
Time scaling: Stretching or compressing time series signals on the timeline can increase the diversity of the data set.
Data truncation: Cutting off part of the time series signal can increase the diversity of the data set.
Data rotation: Rotating time series signals by a certain angle in space can increase the diversity of the data set.
Data flipping: flipping time series signals spatially can increase the diversity of the data set
原文链接：https://blog.csdn.net/qq_42580947/article/details/130156688

6.
(i) Convolution-based models:
In convolutional models, the window size is usually related to the size of the convolution kernel. During the inference phase, the window size determines how the model extracts features from the input data. A larger window can capture broader contextual information, but may increase computational complexity. Convolutional layers are usually designed to handle inputs of different lengths by adjusting padding and stride to accommodate windows of different sizes.

(ii) Recurrent-based models:
When a recurrent neural network (such as LSTM or GRU) processes sequence data, the window size defines the number of features input at each time step. At inference time, recurrent models are able to handle variable-length input sequences, so the window size can be flexibly set to match the needs of specific tasks. Models typically accept inputs of arbitrary length, but when processing very long sequences, batching may be required to avoid memory overflow.

(iii) Transformer-based models:
The transformer model relies on a self-attention mechanism to process sequences, where the window size is usually embodied as an attention window. This architecture does not require step-by-step processing of data like a recurrent neural network, so the entire sequence can be processed in parallel. During inference, you can limit the attention range of the model by setting a fixed window size, or use global attention to cover the entire input sequence. For particularly long sequences, blocking or sliding window techniques can be used to keep the computation manageable.
Ref:ChatGPT