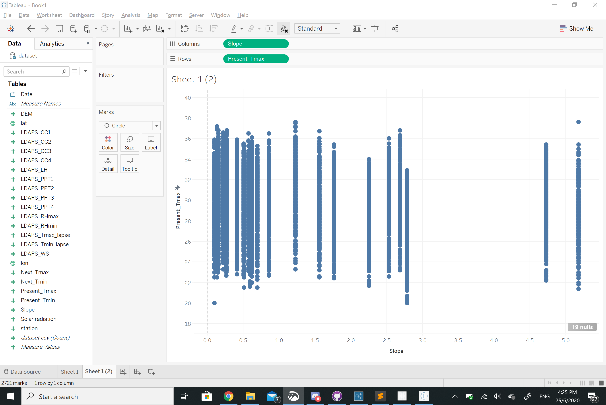
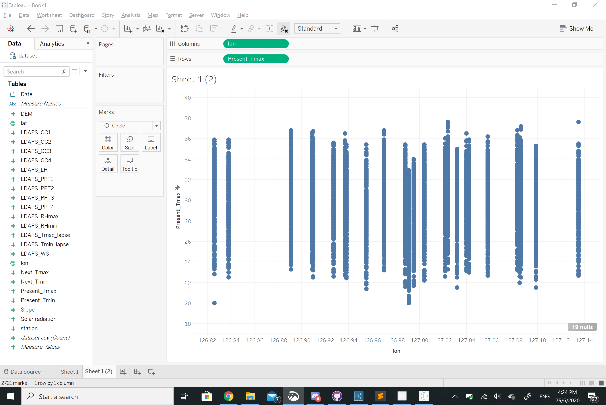
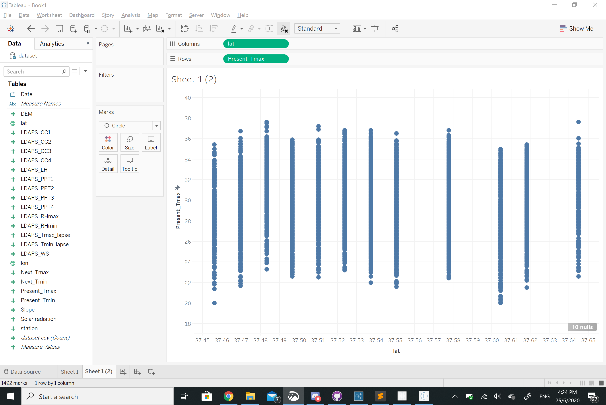
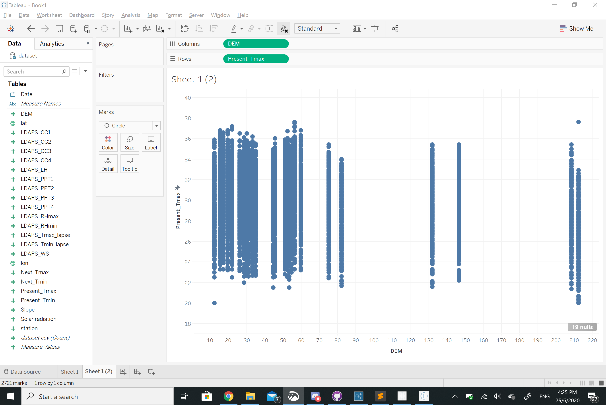
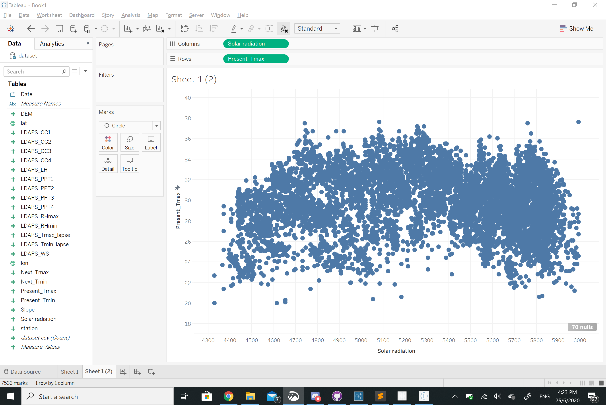
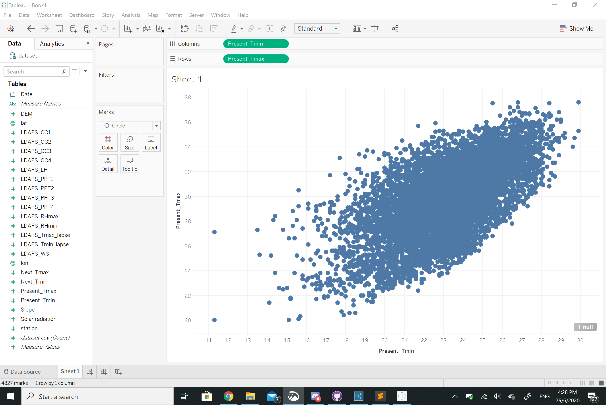
Data Mining Task - Prediction

Data Pre-processing

To predict the next day’s maximum and minimum temperature, we need prepare a prediction model. The first effort of data pre-processing is to prepare a cleansed dataset that could use to train the prediction model.

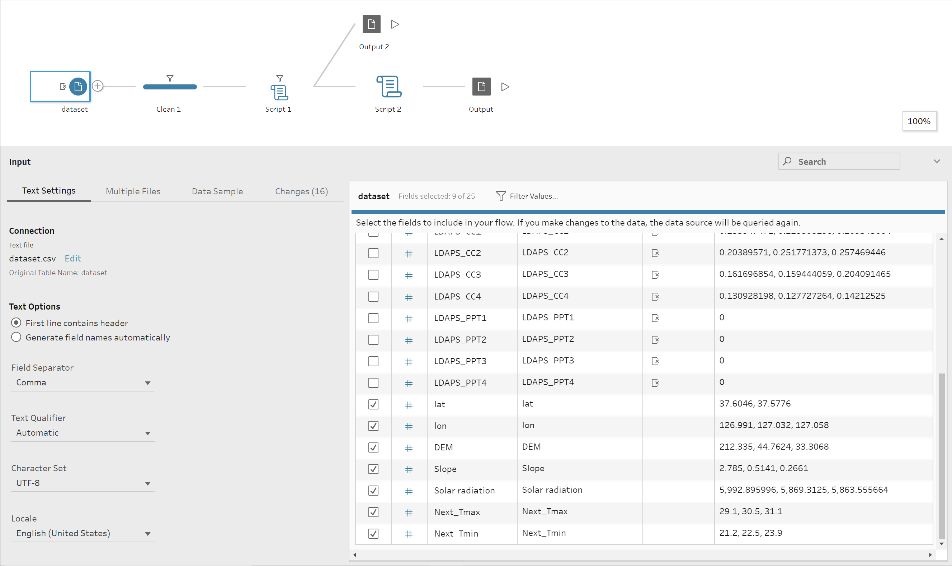
1. Identify influencing factors



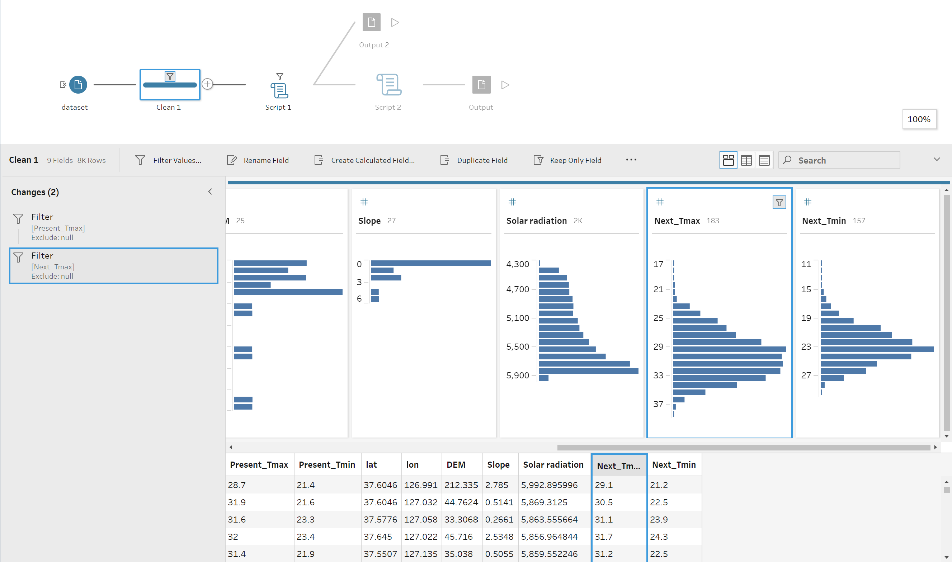
By using Tableau Desktop, we identified the influencing factor by plotting graphs.

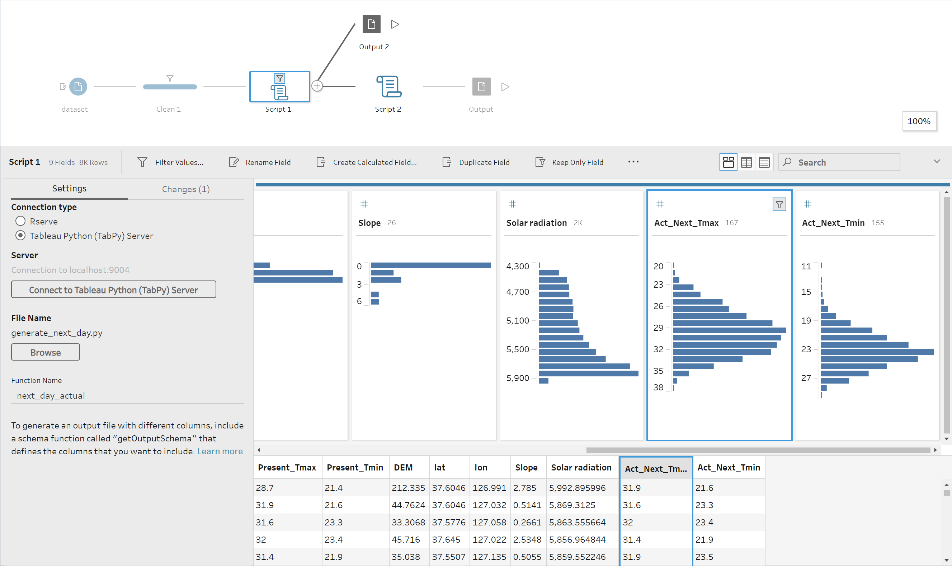
Since we are interested in maximum and minimum temperature, the temperature data is placed at the X axis, and other data are placed at the Y axis. Observing how the data is distributed, we can tell that the relationship between these data. From the graphs, we can tell relationship between solar radiation and maximum temperature, as well as relationship between minimum temperature and maximum temperature resembles a linear pattern. Hence, we resolve to use liner regression model for our prediction model. Latitude, longitude, slope and elevation can easily to be classified as constants, as these data remains the same for each of the observation station. These constants help to increase precision of the model.

1. Prepare the cleansed dataset

Dataset cleansing is done via Tableau Prep Builder. 

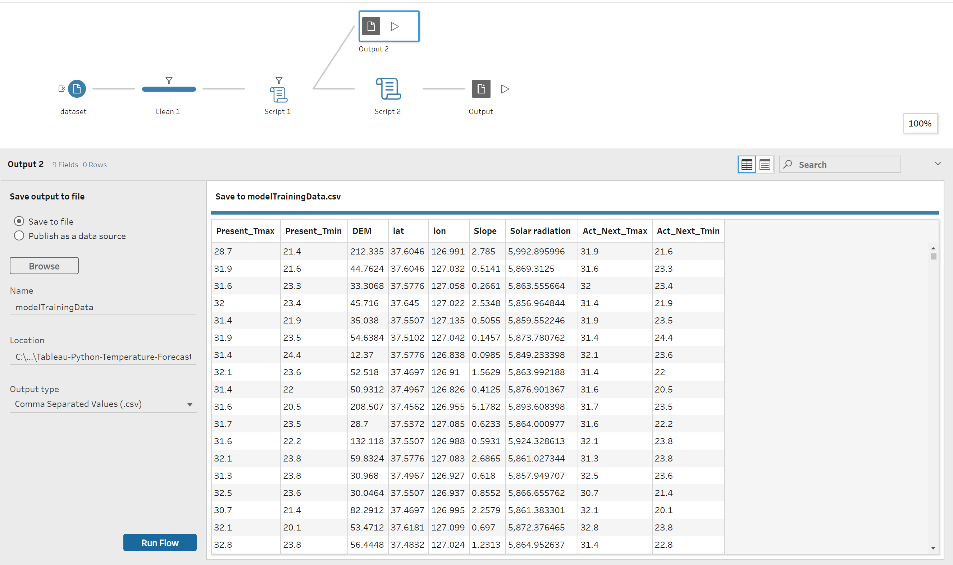
From the input dataset, we include the influencing factor that has been identified previously.



Next, we exclude record that include null value.

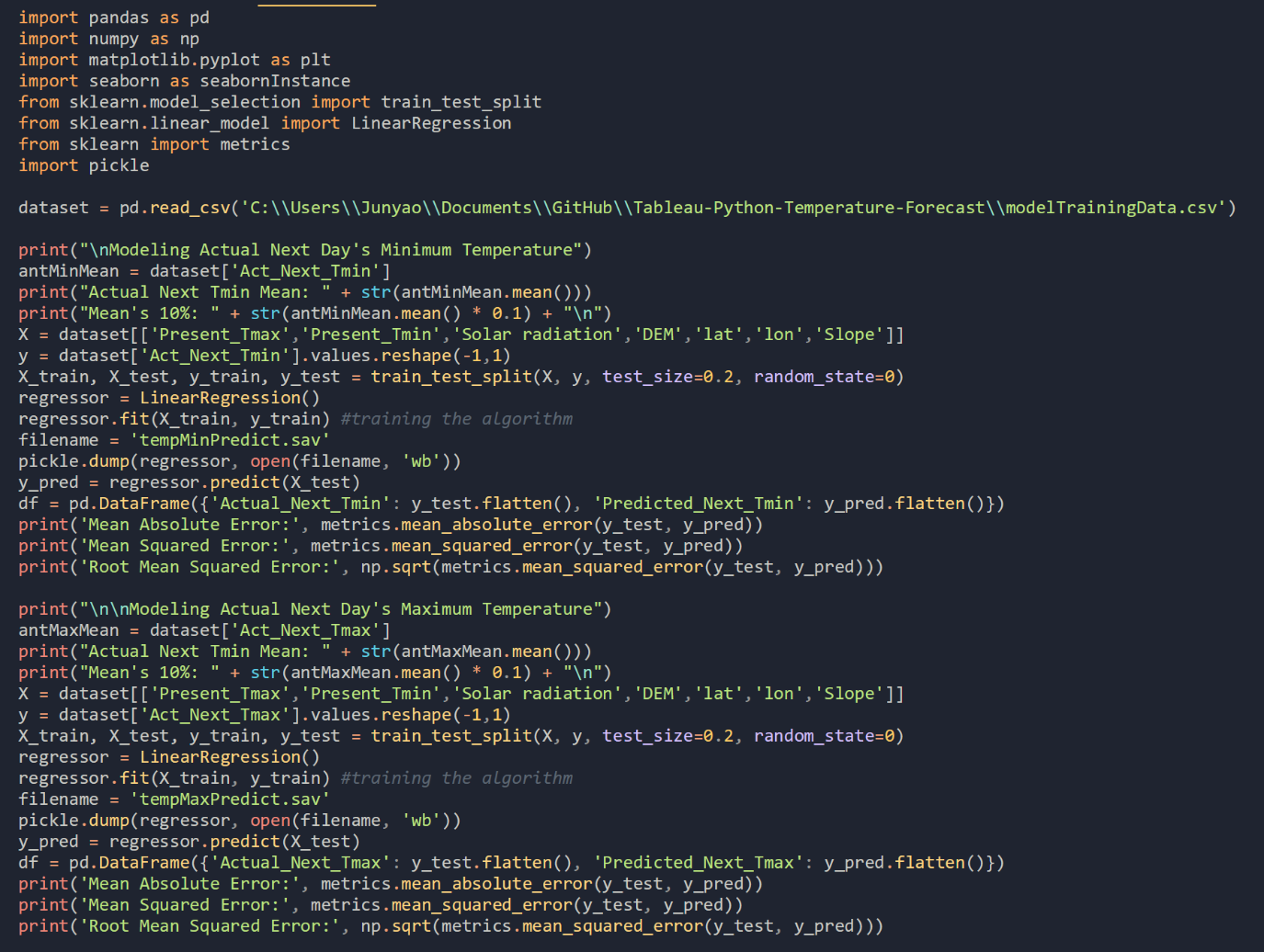
To train the model, we need to provide the actual temperature of the next day. This is achieved using TabPy API, which is an API that allows data to be processed using Python script. With proper setup of local TabPy server, the data is processed using the following Python script

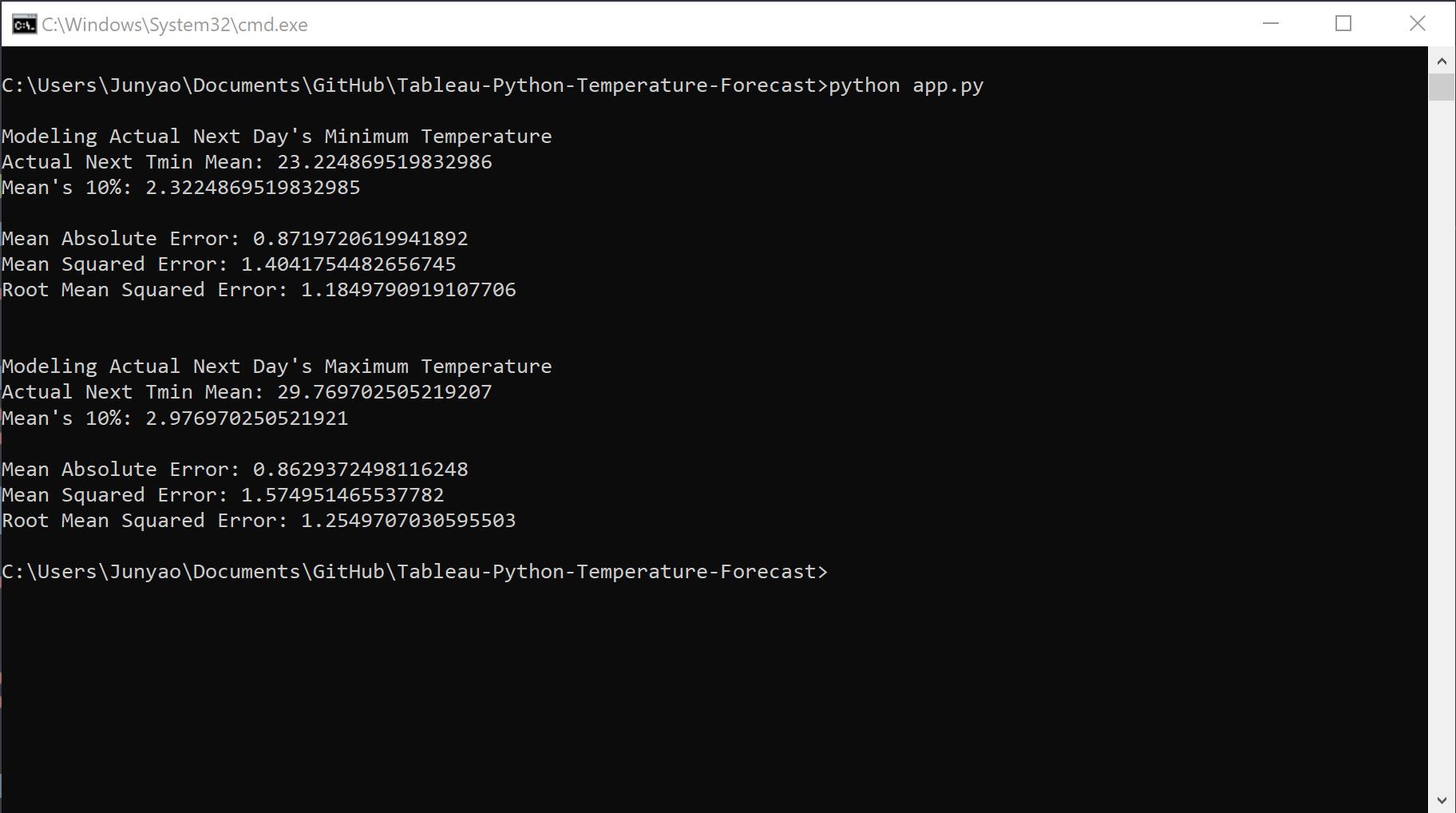


We generate 2 new columns of data, named ‘Act\_Next\_Tmax’ and ‘Act\_Next\_Tmin’, which represents the actual maximum and minimum temperature respectively. The Python script assigns the actual maximum and minimum temperature by getting the next row’s present day maximum and minimum value, which is achieved using the shift function from the pandas library (A Python Library).

The processed data is then being exported as csv file format as input for prediction model training.

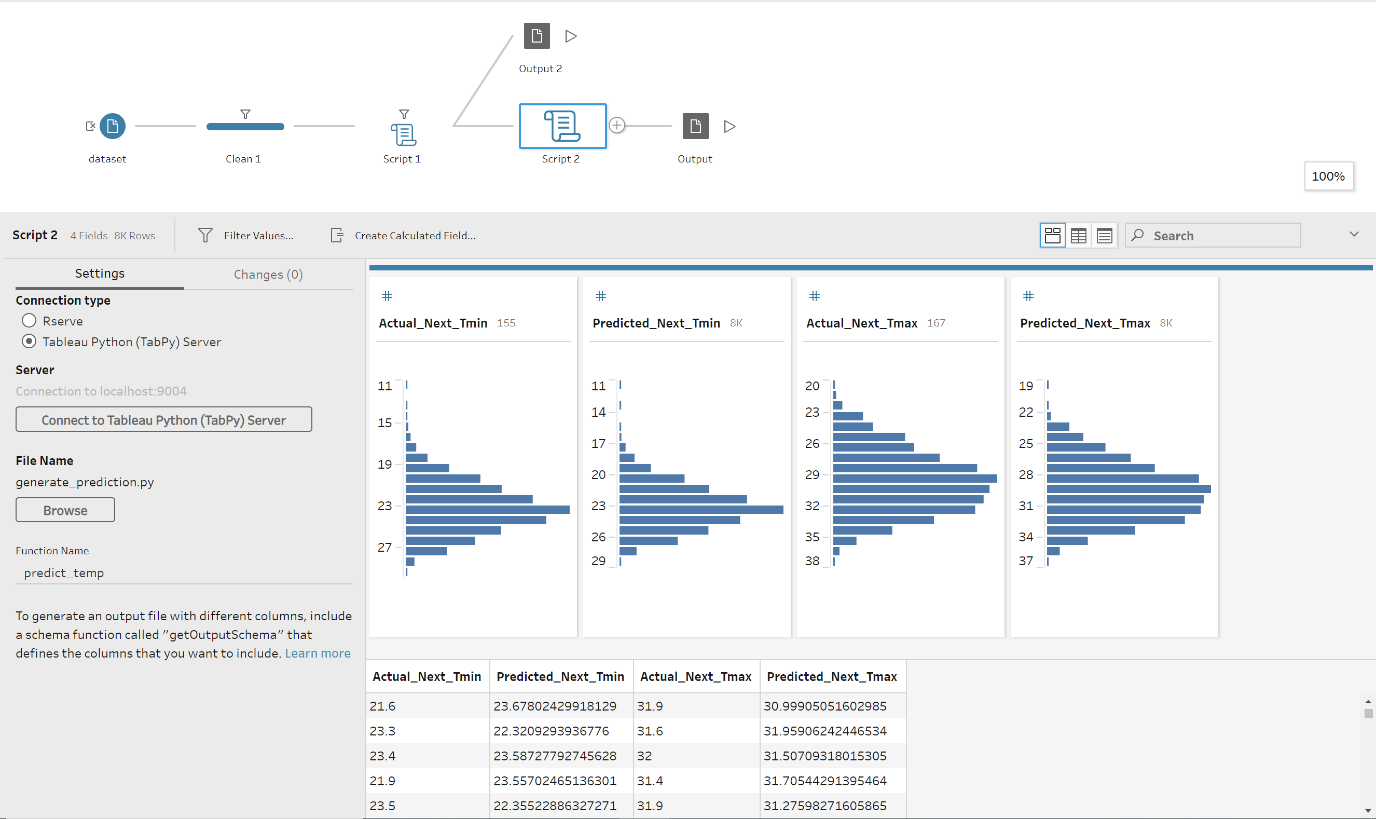
1. Train the prediction model

Once we obtained the cleansed dataset, we can then train our model to predict the next day’s maximum temperature and minimum temperature. We use the Python machine learning library, scikit-learn, to create our prediction model. As we discussed above, we will be using a linear-regression model. The Python snippet for modelling is as below:

We will dump our model into 2 save files, one for maximum temperature and minimum temperature.

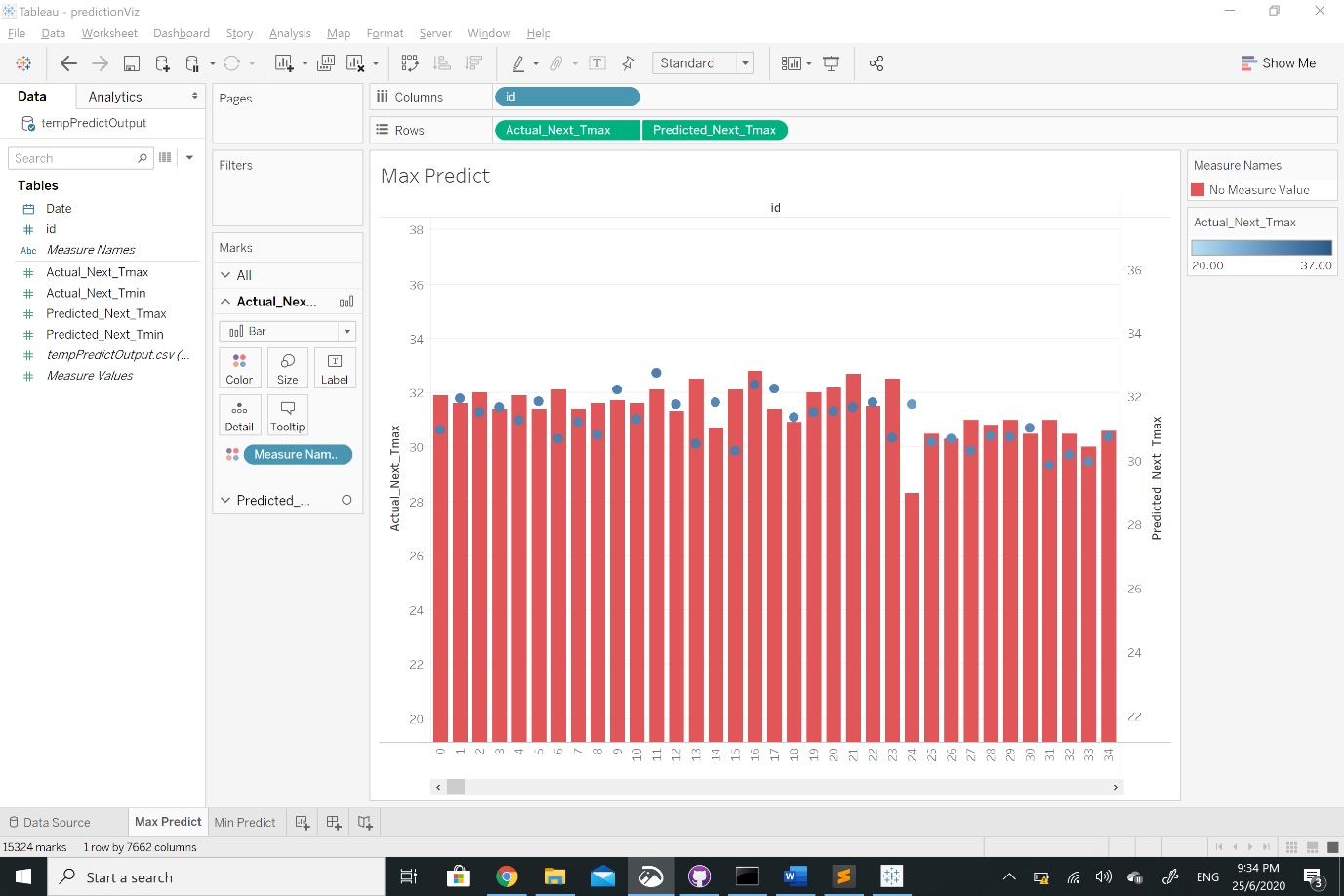
In order to evaluate whether our prediction model is accurate enough, we calculated the Root Mean Squared Error for both maximum and minimum temperature prediction. Gratefully, both maximum and minimum temperature prediction’s Root Mean Squared error is smaller than 10% mean for both maximum and minimum temperature. Hence our prediction model is accepted.

Data Prediction

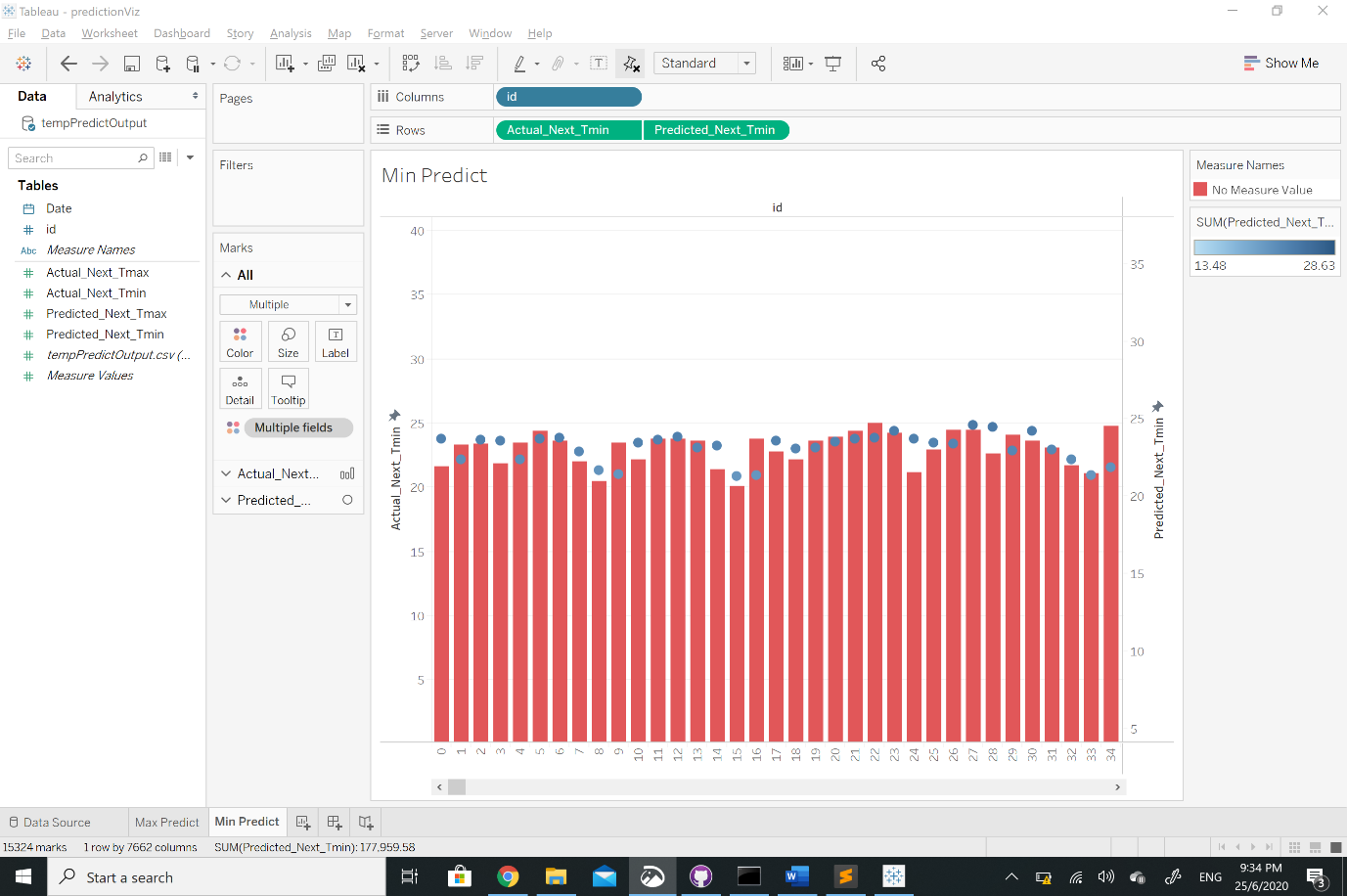
Once we had prepared our prediction model through a series of pre-processing effort, we can now predict the next day’s maximum and minimum temperature.

In Tableau Prep Builder, we once again make use of TabPy API, where we channel our data into the Python script which will then return a dataframe that will contain the prediction value. As in our case, we included the actual temperature for comparison, however in an actual use case where actual temperature is absent or unknown, our model can be modified to return only predicted results.

Finally, the return result is being written into a CSV file for visualisation in Tableau Desktop. The red bars is the actual temperature, where the blue dots represent the predicted value.



Maximum Prediction Visualization



Minimum Prediction Visualization

Discussion

We observe that the prediction is not 100 percent accurate but have a satisfying performance where difference between the actual value and predicted value did not differ too much.

The mean absolute error of our model is 0.8629, which means the average of the difference between the actual and predicted temperature is 0.8629 degree Celsius. In order to increase the accuracy of the model, in the future we can discover and include more influencing factors and also try to use other models to experiment which can yield a better model.