Analyzing Daily Max Temperature:HW5

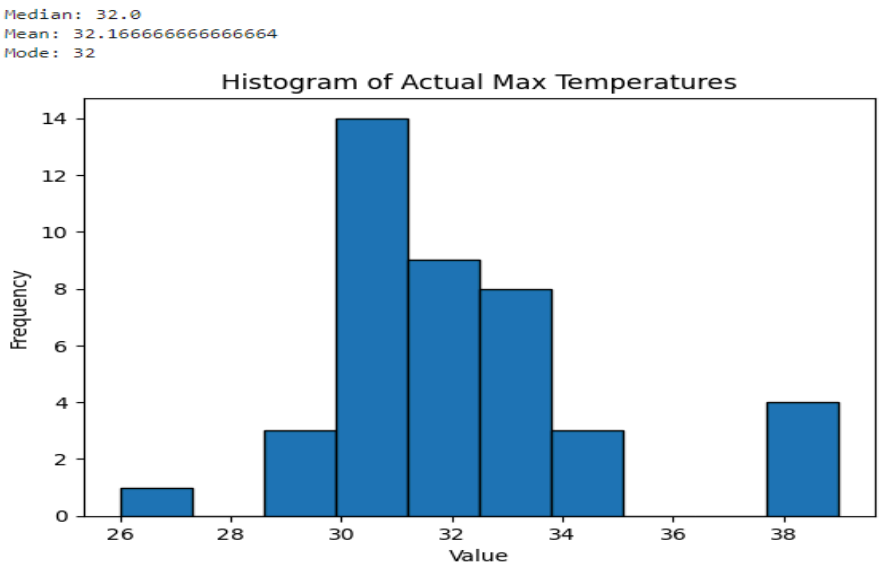
Under Supervision of
Dr. Akansha Rathore

Course Title: Foundations of Data Science

Part 1: **Data Collection:** The csv file named raw_data has already been attached.

Part 2: **Data Cleaning**: I had two missing values on 8th of June and 19th of June. I have used the average of the last three days' predicted temperatures to impute the values.I expect that expected value of predicted temperature of the last three days to be equal to actual temperature of that day. Coming to the point where we have missing value of the next three days' predicted temperature on 8th June and 19th June, I have used the same formula and calculated the predicted temperatures.

Part 2: **Exploratory Statistics**: I have used pandas and matplotlib libraries to calculate mean,median and mode and to plot the histogram.Below is the image.



Part 2: **Processed Data:**The raw data has been processed and updated.It has also been attached.

Part 3: **Model Training and Prediction:**I trained two models using the first 3 and 27 data points respectively.I got the following results.We expect the second model with more data points to be more accurate.I have used scikit-learn library.

[14]:
```
import numpy as np
from sklearn.linear_model import LinearRegression

actual_temps = dataset['Actual_Max_Temperature'].values

X_train1 = np.arange(3).reshape(-1, 1)
y_train1 = actual_temps[:3]
model1 = LinearRegression().fit(X_train1, y_train1)
X_pred1 = np.arange(3, 6).reshape(-1, 1)
predictions1 = model1.predict(X_pred1)

X_train2 = np.arange(27).reshape(-1, 1)
y_train2 = actual_temps[:27]
model2 = LinearRegression().fit(X_train2, y_train2)
X_pred2 = np.arange(27, 30).reshape(-1, 1)
predictions2 = model2.predict(X_pred2)

predictions1, predictions2
```

[14]:
```
(array([39., 39., 39.]), array([29.30769231, 29.04395604, 28.78021978]))
```

Part 3: **Error Calculation and Plotting:**I have used the mean absolute errors to calculate errors for model1 and model2 and app's 3 predictions.Each day has 3 sets of predictions. Here are my results.Also,the errors have been plotted.

```python
from sklearn.metrics import mean_absolute_error

app_predictions = dataset[['Predicted_Temperature_Day1', 'Predicted_Temperature_Day2', 'Predicted_Temperature_Day3']].values

actual_next_3_days = actual_temps[3:6]
app_predictions_next_3_days = app_predictions[:3, :]

errors_model1 = [mean_absolute_error([actual_next_3_days[i]], [predictions1[i]]) for i in range(3)]
errors_model2 = [mean_absolute_error([actual_next_3_days[i]], [predictions2[i]]) for i in range(3)]

errors_app = [mean_absolute_error([actual_next_3_days[i]], [app_predictions_next_3_days[i, j]]) for i in range(3) for j in range(3)]

errors_app_lag1 = errors_app[0::3]
errors_app_lag2 = errors_app[1::3]
errors_app_lag3 = errors_app[2::3]

mean_error_model1 = np.mean(errors_model1)
mean_error_model2 = np.mean(errors_model2)
mean_error_app_lag1 = np.mean(errors_app_lag1)
mean_error_app_lag2 = np.mean(errors_app_lag2)
mean_error_app_lag3 = np.mean(errors_app_lag3)

mean_error_model1, mean_error_model2, mean_error_app_lag1, mean_error_app_lag2, mean_error_app_lag3
```

```
(3.6666666666666665,
 6.289377289377289,
 3.3333333333333335,
 2.3333333333333335,
 2.3333333333333335)
```