Week8 Assignment - Time Series Modeling

Author: Anjani Bonda

Date: 10/22/2022

You will be using the dataset us_retail_sales.csv for this assignment. This data gives the total monthly retail sales in the US from January 1992 until June 2021. With this dataset, complete the following steps:

In [1]:

```
## Importing libraries required for this assignment
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from datetime import datetime
```

In [2]:

```
## Display all columns in pandas dataframe
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

Load the Sales dataset.

In [3]:

```
## Load the data into a dataframe
retail_df = pd.read_csv('us_retail_sales.csv')
retail_df.head(5)
```

Out[3]:

	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	
0	1992	146925	147223	146805	148032	149010	149800	150761.0	151067.0	152588.0	15
1	1993	157555	156266	154752	158979	160605	160127	162816.0	162506.0	163258.0	16
2	1994	167518	169649	172766	173106	172329	174241	174781.0	177295.0	178787.0	18
3	1995	182413	179488	181013	181686	183536	186081	185431.0	186806.0	187366.0	18
4	1996	189135	192266	194029	194744	196205	196136	196187.0	196218.0	198859.0	20

In [4]:

```
## Printing number of rows and columns of the loaded dataframe
retail_df.shape
```

Out[4]:

```
(30, 13)
```

In [5]:

```
## Printing the dtype for each of the column
retail_df.dtypes
```

Out[5]:

```
YEAR      int64
JAN       int64
FEB       int64
MAR       int64
APR       int64
MAY       int64
JUN       int64
JUL     float64
AUG     float64
SEP     float64
OCT     float64
NOV     float64
DEC     float64
dtype: object
```

In [6]:

```
## Looking at summary information about your data (total, mean, min, max, freq,
retail_df.describe()
```

Out[6]:

	YEAR	JAN	FEB	MAR	APR	
count	30.000000	30.000000	30.000000	30.000000	30.000000	30.0
mean	2006.500000	304803.833333	305200.900000	307533.566667	306719.600000	309205.6
std	8.803408	97687.399232	96682.043053	100002.422696	98207.161171	99541.
min	1992.000000	146925.000000	147223.000000	146805.000000	148032.000000	149010.0
25%	1999.250000	228856.750000	231470.750000	233019.000000	233235.500000	234976.5
50%	2006.500000	303486.000000	304592.500000	308655.500000	311233.500000	308690.0
75%	2013.750000	371527.000000	377008.500000	379221.000000	376797.500000	382698.2
max	2021.000000	520162.000000	504458.000000	559871.000000	562269.000000	548987.0

Perform Exploratory Data Analysis

In [7]:

```
# Use melt to convert from wide to long format
retail_df2 = pd.melt(retail_df, id_vars='YEAR', value_vars=['JAN', 'FEB', 'MAR'
                                                             'APR', 'MAY', 'JUN', 'JUL', 'AUG'
```

In [8]:

```
# Convert Year to string
retail_df2['YEAR'] = retail_df2['YEAR'].astype(str)
```

In [9]:

```
# Build a new column for date
retail_df2['Date'] = retail_df2['variable'] + '-01-' + retail_df2['YEAR']

# Convert Date to Datetime
retail_df2['Date'] = pd.to_datetime(retail_df2['Date'])
```

In [10]:

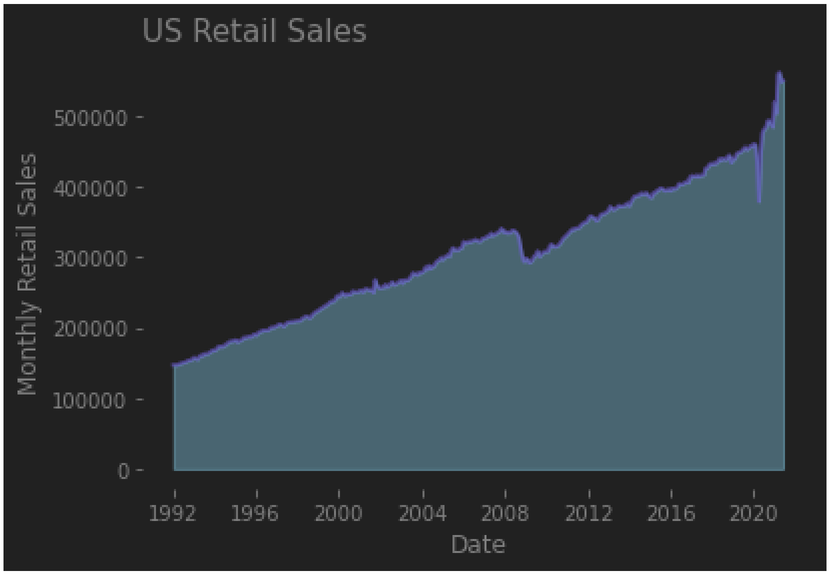
```
# Drop NA
retail_df2.dropna(inplace=True)
```

In [11]:

```
# Sort by date
retail_df2 = retail_df2.sort_values(by=['Date'])
```

1. Plot the data with proper labeling and make some observations on the graph.

```
In [12]: # Create an area chart
         plt.fill_between(retail_df2['Date'], retail_df2['value'], color="skyblue", alph
         plt.plot(retail_df2['Date'], retail_df2['value'], color="Slateblue", alpha=0.6,
         plt.box(False)
         plt.title('US Retail Sales', loc='left', fontsize=15, color='grey')
         plt.xlabel('Date', fontsize=12, color='grey')
         plt.ylabel('Monthly Retail Sales', fontsize=12, color='grey')
         plt.tick_params(axis='x', colors='grey')
         plt.tick_params(axis='y', colors='grey')
         plt.show()
```



Observation

US Retail sales have been steadily increasing since 1992. As you can see in the chart, small decreases in retail sales were seen during the housing crisis (2008-2009) and at the beggining of the pandemic (2020)

2. Split this data into a training and test set. Use the last year of data (July 2020 – June 2021) of data as your test set and the rest as your training set.

```
In [13]: # Build a new feature from date to be used as a predictor (using ordinal time)
         retail_df2['O-Date'] = pd.to_datetime(retail_df2['Date'])
         retail_df2['O-Date'] = retail_df2['O-Date'].map(datetime.toordinal)
```

```
In [14]: # Build a new predictor for month
         months = dict(JAN=1, FEB=2, MAR=3, APR=4, MAY=5, JUN=6, JUL=7, AUG=8, SEP=9, OC
         retail_df2['Month'] = retail_df2['variable'].map(months)
```

```
In [15]: ## Spliting based on row value
         training = retail_df2.iloc[0:341]
         test = retail_df2.iloc[342:354]
```

```
In [16]: # Split out x & y reshape date fields
         x_train = training[['O-Date', 'Month']]
         y_train = training['value']
```

```
x_test = test[['O-Date', 'Month']]
y_test = test['value']
```

3. Use the training set to build a predictive model for the monthly retail sales.

In [17]:

```
# Create a model
model = LinearRegression()

# Fit the model to the training set
model.fit(x_train, y_train)
```

Out[17]:

```
▾ LinearRegression
LinearRegression()
```

4. Use the model to predict the monthly retail sales on the last year of data.

In [18]:

```
# Predict the last years retail sales
test_predictions = model.predict(x_test)
```

5. Report the RMSE of the model predictions on the test set.

In [19]:

```
print('Test RMSE:', metrics.mean_squared_error(y_test, test_predictions, square
```

```
Test RMSE: 66817.27313121158
```

A large spike in retail sales was seen during the period of time the model is attempting to predict. This is likely causing the increased RMSE.

In [ ]: