Amazon Holiday Sales Analysis

Cara Hsiao

Agenda

Project Intro
Main Idea, Purpose, & Focus

Dataset Overview
Contents, Origin, Size, Source

Data Exploration
Visualizations, Analysis, & Findings

Limitations
Weaknesses & Challenges

Business Insights
Key Takeaways & Real Word Value

Project Goals

Observe sales trends surrounding the holiday season

Identify what factors influence holiday sales figures

Create model to predict sales figures given key factors

Dataset Introduction

Background

Sales data for 45 Amazon stores in different regions

Creator

Published by Revanth Krishna on Kaggle

Date Range

Data Range from Jaunuary 10th 2019 - December 10th 2021, average of 7.5 days between measurements

Size

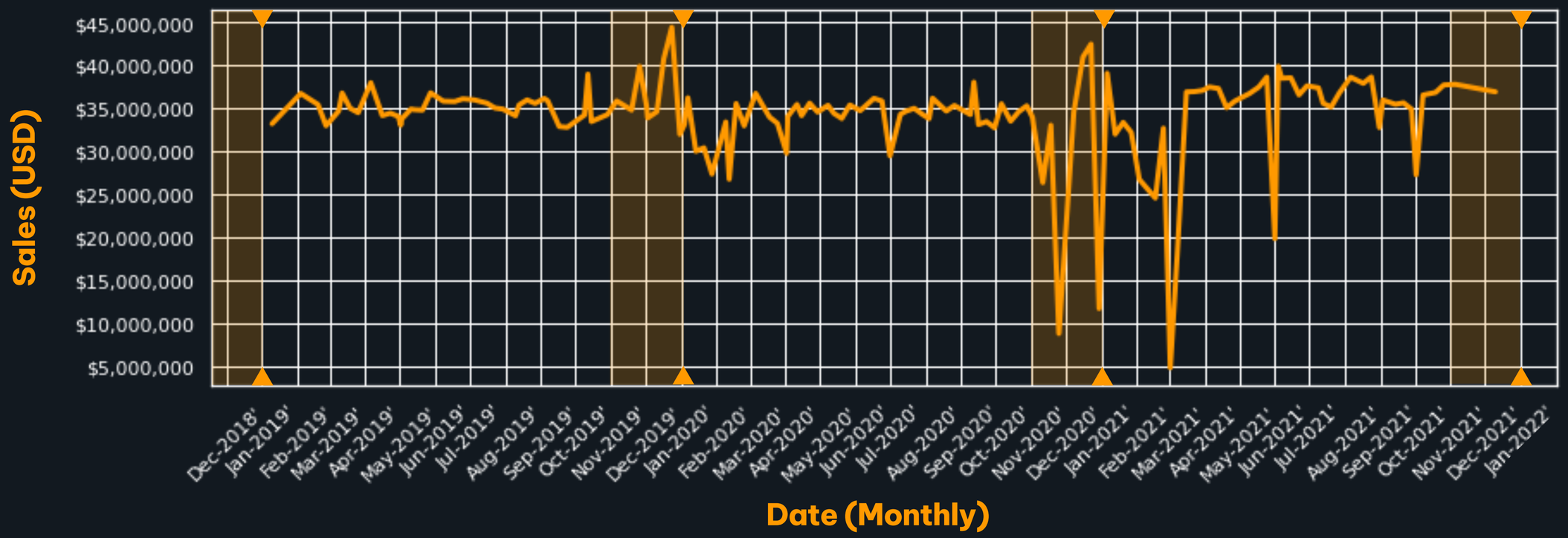
374,247
ROWS

21
COLUMNS

Data Visualization

Weekly Aggregated Store Sales

(*Holidays Highlighted Orange, Years Delineated)



Introduction | Dataset | Exploration | Limitations | Business Insight

Pre-Processing

- Separated Date into Day, Month, Year columns
- Created mask for months <10 for November and December to isolate holidays
- Removed unexplained/ unnecessary features

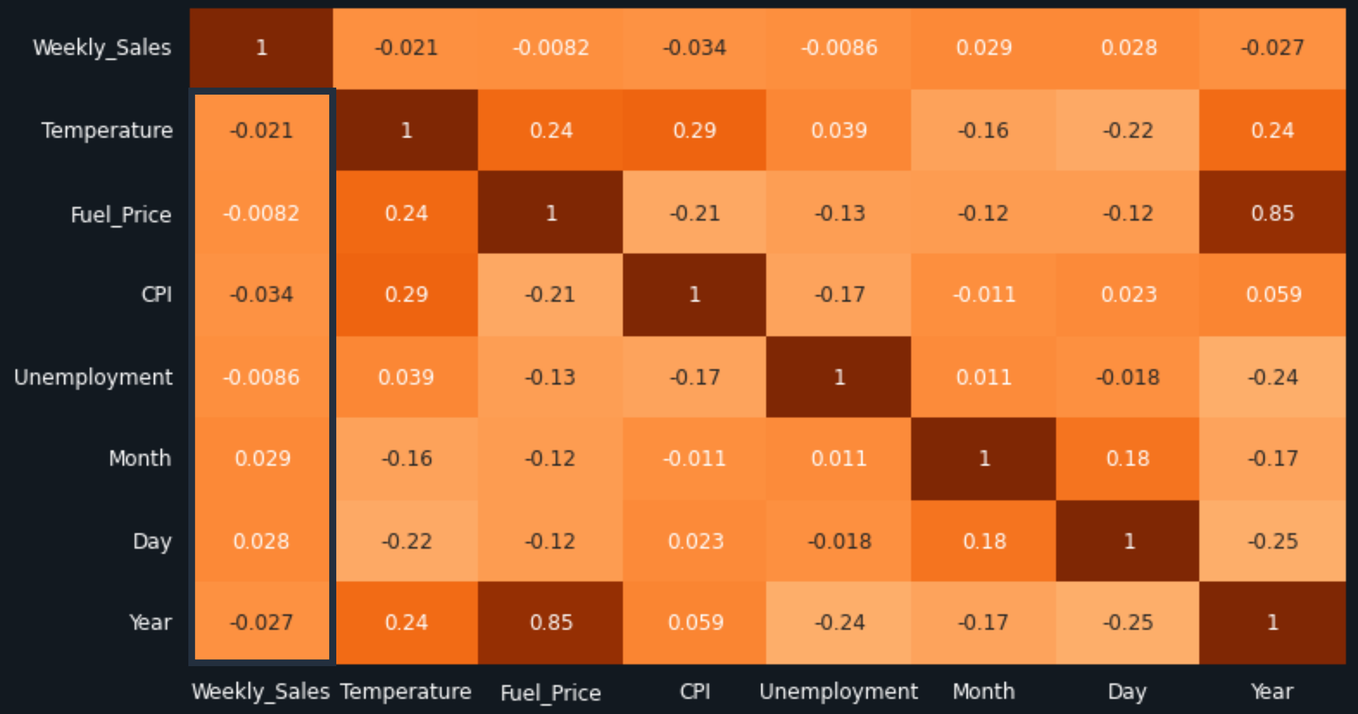
Remaining Features

(Dataset contains 49192 rows and 8 columns)

- Temperature
- Fuel Price
- CPI
- Unemployment
- Day
- Month
- Year
- Weekly Sales

Holiday Correlation Heatmap

November & December



Observations

- Weak correlations (<5%) between features/sales
- Strongest correlation to month and day
- Expected negative correlations with, fuel price, temperature, CPI and unemployment

Multiple Linear Regression

	Corr.	P-val.
Temperature	-.08619	0.986
Fuel Price	913.58	0.029
CPI	-12.516	0.000
Unemployment	-258.33	0.000
Month	686.93	0.000
Day	27.916	0.001
Year	-985.45	0.000

Observations

- Fuel Price, Month, and day had positive correlations to sales
- Temperature, CPI, Unemployment, and Year had negative correlations to sales
- Sales increased towards the end of the month and the end of the year
- Temperature had a P value >.05 indicating that it is a poor predictor of weekly sales

Multiple Linear Regression

Results

R^2: 0.003274211934140 90
MSE: 247497900.2818167

Weekly Sales

1,996,819-.8619(Temp)+913.58(Fuel)-258.33(Unemp.)
-12.516(CPI)+686.93(Month)+27.916(Day)-686.93(Year)

Model

Training data size: (36894, 7)
Testing data size: (12298, 7)

Training set RMSE: 15806.64
Test set RMSE: 15507.68

Observations

- R^2 value is .327% indicating the model's fit is poor
- This model cannot produce accurate sales predictions
- Large MSE indicates mean level of variance is high, meaning further indicating model inaccuracy

Limitations

- Key feature markdown was incomplete
- Data spanned too short a time period
- Data lacked important predictive factors
- Time-span between dates was large and inconsistent
- Methods used to analyze dataset are not ideal
- Data was from 45 storefronts out of 30,000

Business Insights

Key Takeaways

1. Sales increase higher sales towards the month's end

2. High unemployment and inflation reduce sales

Seller Insight

Model can inform timings for:

- Advertising
- Promotions
- Restocking

Amazon Insight

Model can inform timings for:

- Fulfillment center traffic
- Seasonal worker hiring
- Shipping times

Thank you.

Questions?

Glossary

P value: probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct, P values >.05 is insignificant

MSE: measures error in statistical models by using the average squared difference between observed and predicted values.

```
Intercept:
 1996819.3959801344
Coefficients:
 [-8.61854533e-02  9.13584470e+02 -1.25161729e+01 -2.58332283e+02
  6.86926740e+02  2.79163444e+01 -9.85453810e+02]
                            OLS Regression Results
==============================================================================
Dep. Variable:           Weekly_Sales   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     23.08
Date:                Thu, 08 Dec 2022   Prob (F-statistic):           1.70e-31
Time:                        03:29:17   Log-Likelihood:            -5.4517e+05
No. Observations:               49192   AIC:                         1.090e+06
Df Residuals:                   49184   BIC:                         1.090e+06
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.997e+06   5.01e+05      3.984      0.000    1.01e+06    2.98e+06
Temperature      -0.0862      4.897     -0.018      0.986      -9.685       9.512
Fuel_Price      913.5845    418.705      2.182      0.029      92.917    1734.252
CPI             -12.5162      2.356     -5.314      0.000     -17.133      -7.899
Unemployment   -258.3323     58.450     -4.420      0.000    -372.896    -143.769
Month           686.9267    147.757      4.649      0.000     397.321     976.533
Day              27.9163      8.620      3.239      0.001      11.022      44.811
Year           -985.4538    248.680     -3.963      0.000   -1472.870    -498.037
==============================================================================
Omnibus:                    13416.613   Durbin-Watson:                   1.971
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            29556.399
Skew:                           1.595   Prob(JB):                         0.00
Kurtosis:                       5.062   Cond. No.                     1.43e+07
==============================================================================
```

import pandas as pd
import numpy as np
from scipy import stats
import math
from sklearn import datasets
import statsmodels.api as sm

```python
import pandas as pd
from sklearn import linear_model
import statsmodels.api as sm


x = holiday_data[feature_list]
y= holiday_data.Weekly_Sales

# with sklearn
regr = linear_model.LinearRegression()
regr.fit(x, y)

print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)

# with statsmodels
x = sm.add_constant(x) # adding a constant

model = sm.OLS(y, x).fit()
predictions = model.predict(x)

print_model = model.summary()
print(print_model)
```

```
[13] holiday_data = holiday_data.drop('Date', axis=1)
```

```
[14] shape=holiday_data.shape
     print("Dataset contains {} rows and {} columns".format(shape[0],shape[1]))
```

```
Dataset contains 49192 rows and 8 columns
```

```
[15] holiday_data.head (5)
```

	Weekly_Sales	Temperature	Fuel_Price	CPI	Unemployment	Month	Day	Year
42103	77.38000	80.84000	2.66800	126.11190	9.59300	11	6	2019
42104	21800.56000	78.45000	2.66800	126.11190	7.89600	11	6	2019
42105	28996.47000	61.64000	2.97200	132.43574	8.18500	11	6	2019
42106	24454.43000	68.90000	2.80900	136.28743	9.05100	11	6	2019
42107	1852.76000	83.75000	2.66800	215.01872	6.38400	11	6	2019

```
[16]
     feature_list = ['Temperature', 'Fuel_Price', 'CPI', 'Unemployment', 'Month', 'Day', 'Year']
     X_mult = holiday_data[feature_list]
     y_mult = holiday_data.Weekly_Sales

     # instantiate and fit like last time
     mult_linreg = LinearRegression()
     mult_linreg.fit(X_mult, y_mult)

     # print intercept
     print("The y intercept:", mult_linreg.intercept_)
```