CDS6214

Data Science Fundamentals

Project (40%)

Tutorial Section: TT4L

Group Number: G03

ID	Name	Email Address
1211102777	Tai Qi Tong	1211102777@student.mmu.edu.my
1211102630	Chan Kar Kin	1211102630@student.mmu.edu.my

YouTube Link	https://youtu.be/-EMsTZR8H64

TASK DISTRIBUTION

Lecture Section: TC1L

Tutorial Section: TT4L

Group Number: G03

Domain: Retail

Task Distribution:

Student Name	Task Done (Be very specific)
Tai Qi Tong	Question 1 & 2, Data pre-processing, EDA, Data modelling & analysis
Chan Kar Kin	Problem statement, Data pre-processing, Question 3 & 4, Data collection, Data modelling & analysis

Table of Contents

1.0 Introduction

This report is to present our findings from the EDA and data models on the Adidas US Sales Dataset from Kaggle. The data science processes are performed using Python in Jupyter Notebook.

1.1 Motivation

The motivation behind this analysis is to enhance the operational efficiency of adidas in the US market, aiming to achieve greater business agility. By understanding the underlying factors influencing sales and profitability, adidas can make informed decisions to optimize its operations.

1.2 Problem Statement

The goal is to analyze the adidas sales dataset to improve operational efficiency in the US market. This involves understanding how seasonal trends, product pricing, product categories, and regional differences impact sales and profitability.

1.3 Impact

There are several impacts of improving operational efficiency to communities and the nation. It can enhance adidas' market presence in the US by ensuring products are readily available where and when consumers demand them. Furthermore, it can strengthen the competitiveness of adidas within the US market, positioning itself more effectively against competitors. These improvements contribute to a stronger market position, fostering sustainable growth and profitability for adidas in the dynamic US marketplace.

2.0 Problem Definition and Questions

2.1 Main Problem

The main problem is to improve operational efficiency at adidas in the US market to enhance business agility.

2.2 Questions

There are four research questions in this project:

1. How do seasonal trends affect sales and profitability across different product categories?
2. How does the pricing of products influence both sales volume and profitability across different types of products?
3. What is the highest total units sold product category based on gender, and how does the price per unit for these categories vary across different states?
4. Is there a significant difference in total sales across cities between weekdays and weekends?

3.0 Data Collection

3.1 Data Sources

The analyzed dataset is taken from Kaggle.

3.2 Data Description

The dataset includes Adidas sales data in the US from 01/01/2020 to 31/12/2021, covering various retailers, product categories, regions, and pricing information. The following are the descriptions of the columns:

Variable	Type	Comments
Retailer	Categorical	Retailer name
Retailer ID	Numerical	Identifier for retailer
Invoice Date	Date/Time	Date and time of invoice
Region	Categorical	Geographic region where the sales operations occur
State	Categorical	State within the region where sales occurred
City	Categorical	City within the state where sales occurred
Product	Categorical	Type of product sold
Price per Unit	Numerical	Price of each unit of product sold
Units Sold	Numerical	Number of units sold in invoice
Total Sales	Numerical	Total sales amount in each invoice (Price per Unit * Units Sold)
Operating Profit	Numerical	Profit derived from total sales
Operating Margin	Numerical	Profit margin (Operating Profit / Total Sales)
Sales Method	Categorical	Method of sales in each invoice

4.0 Data Preprocessing

The following steps are performed for data preprocessing:

1. Gather data from Kaggle. (Adidas Sales Dataset, 2022)
2. Handle missing values and set header:
   We dropped the first three rows that have only null values and set our header for the dataset.

	Retailer	Retailer ID	Invoice Date	Region	State	City	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Operating Margin	Sales Method
1	Foot Locker	1185732	2020-01-01 00:00:00	Northeast	New York	New York	Men's Street Footwear	50	1200	600000	300000	0.5	In-store
2	Foot Locker	1185732	2020-01-02 00:00:00	Northeast	New York	New York	Men's Athletic Footwear	50	1000	500000	150000	0.3	In-store
3	Foot Locker	1185732	2020-01-03 00:00:00	Northeast	New York	New York	Women's Street Footwear	40	1000	400000	140000	0.35	In-store
4	Foot Locker	1185732	2020-01-04 00:00:00	Northeast	New York	New York	Women's Athletic Footwear	45	850	382500	133875	0.35	In-store
5	Foot Locker	1185732	2020-01-05 00:00:00	Northeast	New York	New York	Men's Apparel	60	900	540000	162000	0.3	In-store

3. Check for duplicated rows:
   No duplicated rows found in the dataset.
4. Data type conversion:
   a. Retailer ID, and Units Sold are converted to integer

b. Invoice Date converted to datetime format, extracted Year, Month, Day, Day of Week and Quarter into new columns from Invoice Date
c. Price per Unit, Total Sales, Operating Profit, and Operating Margin are converted to float rounded to two decimal places
d. Gender from Product is extracted into Gender column
e. Price Change, Sales Volume Change and Profitability Change were calculated, each representing the respective changes over time within specific product groups.

5. Handle inconsistent columns:
   a. Column Retailer ID is dropped (there are 4 retailer ID with 5 retailers)
   b. Column Total Sales is recalculated by multiplying Price per Unit with Units Sold
   c. Column Operating Profit is recalculated by multiplying Total Sales with Operating Margin
6. Eliminate columns not in use:
   a. Column Operating Margin and Sales Method is removed.

Retailer	Invoice Date	Quarter	Year	Month	Day	Day of Week	Region	State	City	Gender	Product	Price per Unit	Units Sold	Total Sales	Operating Profit	Price Change	Sales Volume Change
Foot Locker	2020-01-03	2020Q1	2020	1	3	Friday	Northeast	New York	New York	Women	Street Footwear	40.0	1000	40000.0	14000.00	-10.0	-200.0
Foot Locker	2020-01-04	2020Q1	2020	1	4	Saturday	Northeast	New York	New York	Women	Athletic Footwear	45.0	850	38250.0	13387.50	-5.0	-150.0
Foot Locker	2020-01-06	2020Q1	2020	1	6	Monday	Northeast	New York	New York	Women	Apparel	50.0	1000	50000.0	12500.00	-10.0	100.0
Foot Locker	2020-01-07	2020Q1	2020	1	7	Tuesday	Northeast	New York	New York	Men	Street Footwear	50.0	1250	62500.0	31250.00	10.0	250.0
Foot Locker	2020-01-08	2020Q1	2020	1	8	Wednesday	Northeast	New York	New York	Men	Athletic Footwear	50.0	900	45000.0	13500.00	5.0	50.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Foot Locker	2021-01-24	2021Q1	2021	1	24	Sunday	Northeast	New Hampshire	Manchester	Men	Apparel	50.0	64	3200.0	896.00	-11.0	-80.0
Foot Locker	2021-01-24	2021Q1	2021	1	24	Sunday	Northeast	New Hampshire	Manchester	Women	Apparel	41.0	105	4305.0	1377.60	-9.0	41.0

```
data.shape
```

```
(9645, 19)
```

Cleaned data with 9645 rows and 19 columns.

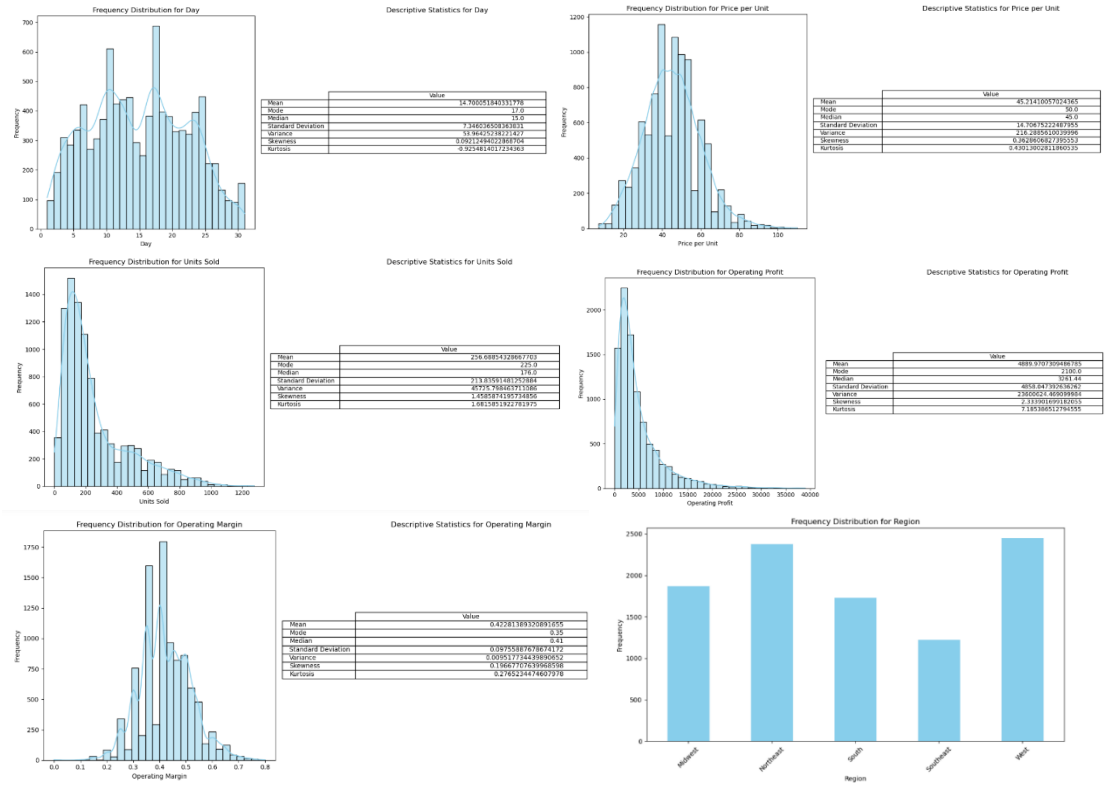
5.0 EDA

1. Distribution of numerical variables and non-numerical variables:
   Average product price is 45.22 USD, average units sold is 256.93 USD, average total sales is 12455.08 USD, average operating profit is 4894.79 USD.

	Retailer	Day of Week	Region	State	City	Gender	Product
count	9645	9645	9645	9645	9645	9645	9645
unique	6	7	5	50	52	2	3
top	Foot Locker	Tuesday	West	California	Portland	Men	Street Footwear
freq	2634	1490	2448	432	360	4823	3217

	Year	Month	Day	Price per Unit	Units Sold	Total Sales	Operating Profit	Price Change	Sales Volume Change	Profitability Change
count	9645.000000	9645.000000	9645.000000	9645.000000	9645.000000	9645.000000	9645.000000	9645.000000	9645.000000	9645.000000
mean	2020.865319	6.459824	14.700052	45.214101	256.688543	12441.954381	4889.970731	-0.004977	-0.294660	-63.114330
std	0.341400	3.453995	7.346037	14.706752	213.835915	12696.347010	4858.047393	10.284423	129.230587	40185.529483
min	2020.000000	1.000000	1.000000	7.000000	0.000000	0.000000	0.000000	-55.000000	-725.000000	-322500.000000
25%	2021.000000	3.000000	9.000000	35.000000	106.000000	4060.000000	1753.050000	-6.000000	-71.000000	-2338.880000
50%	2021.000000	6.000000	15.000000	45.000000	176.000000	7800.000000	3261.440000	0.000000	-2.000000	-17.320000
75%	2021.000000	9.000000	21.000000	55.000000	350.000000	15792.000000	6187.500000	6.000000	72.000000	2259.090000
max	2021.000000	12.000000	31.000000	110.000000	1275.000000	82500.000000	39000.000000	50.000000	550.000000	285750.000000

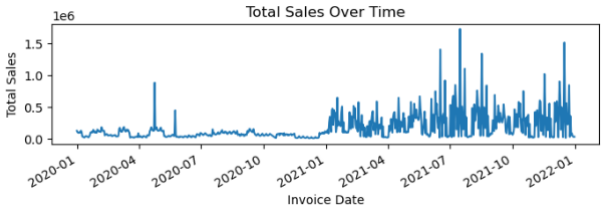
2. Frequency distribution and descriptive statistics for attributes in dataset:



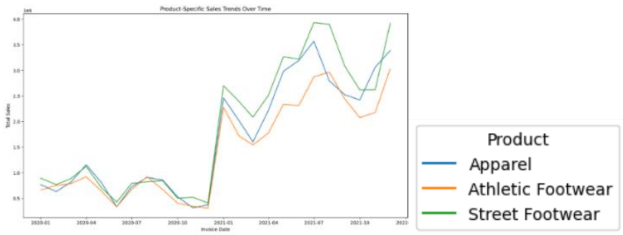
3. Seasonal Sales Pattern

I. Total Sales by Invoice Date

i. The overall sales for 2021 were significantly higher than in 2020, with peak sales occurring around August 2021.

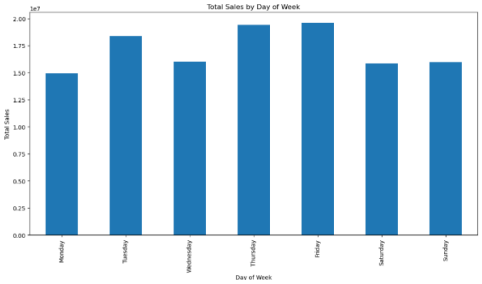


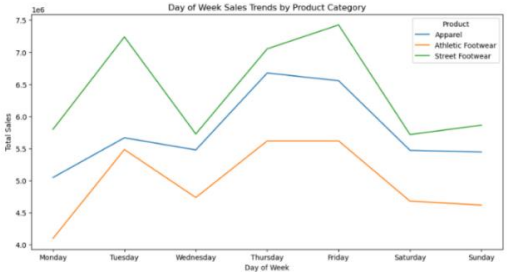
ii. Based on each Product Total Sales Over Time, we found that the sales for each product are similar at year 2020, and slightly different at year 2021. However, most of the time, the product of street footwear shows higher sales than the other two products while athletic footwear is having a lower sales compared to the other two products.



II. Total Sales by Day of Week

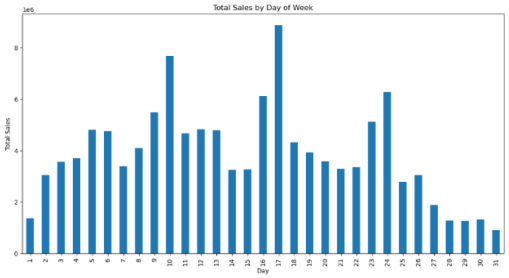
We found that the highest sales of the week were falls on Friday, followed by Tuesday. While Monday is normally having lesser sales compared to other day of week.

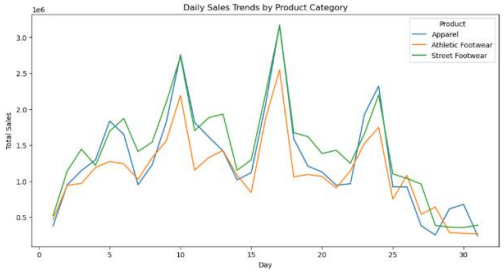




III. Total Sales by Day

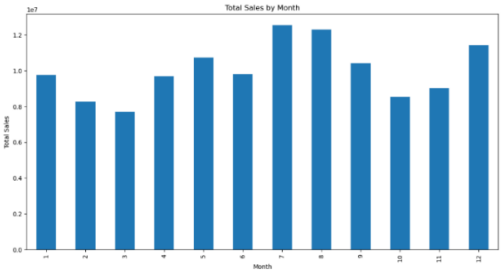
We found that 17th , 10th and 24th of each month have higher sales compared to other days in each month.





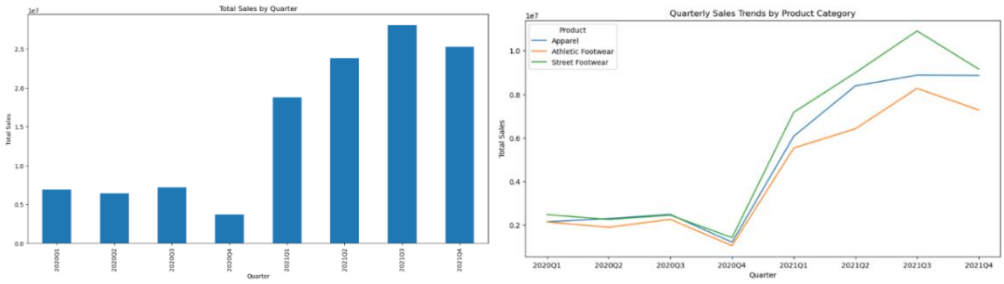
IV. Total Sales by Month

July has the highest sales in each year compared to other months.

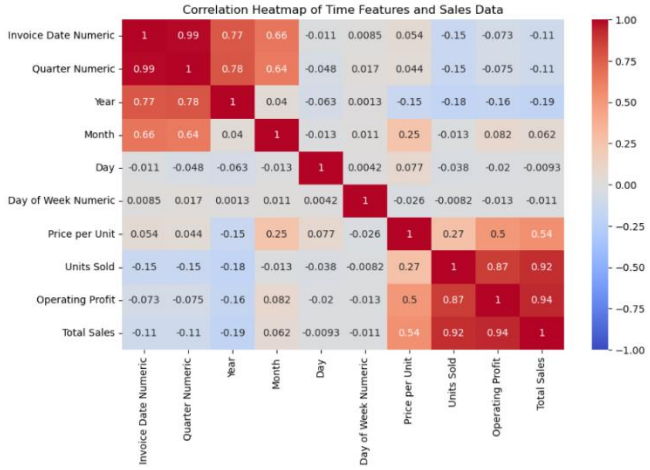


V. Total Sales by Quarter

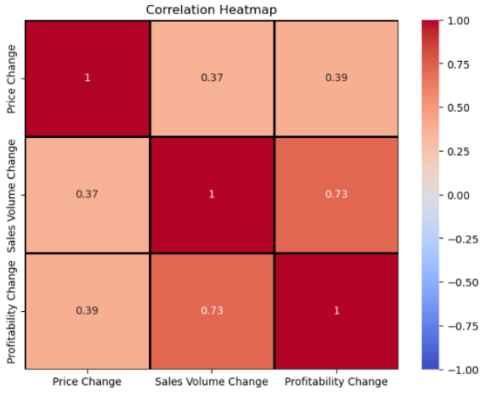
For each quarter observation, quarter 3 of each year has higher sales in the year.



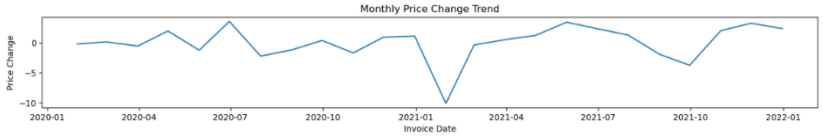
4. Correlation analysis of numerical columns:
   - Units Sold is highly correlated with total sales and operating profit.
   - Total Sales is highly correlated with price per unit, units sold and operating profit.
   - Operating Profit is highly correlated with price per unit, units sold and total sales.



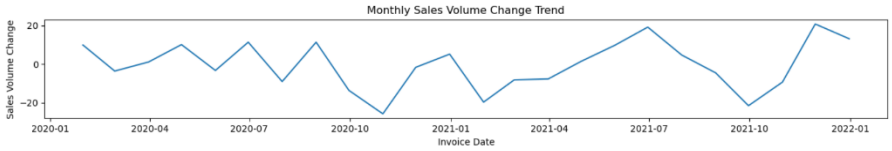
5. Correlation analysis using a heatmap indicates profitability are positively correlated with sales volume change and price change. Notably, an increase in sales volume is strongly associated with a significant increase in profitability, with a correlation coefficient of 0.73.



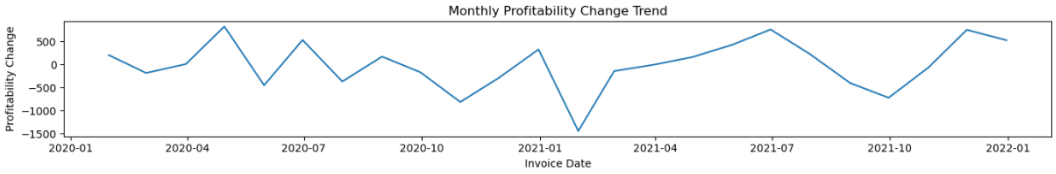
6. Monthly Trends
   I. Monthly Price change trend
   There's a significant price drop around January 2021, followed by a recovery suggesting upward price adjustments. Monthly fluctuations reveal frequent price adjustments with multiple peaks and valleys.



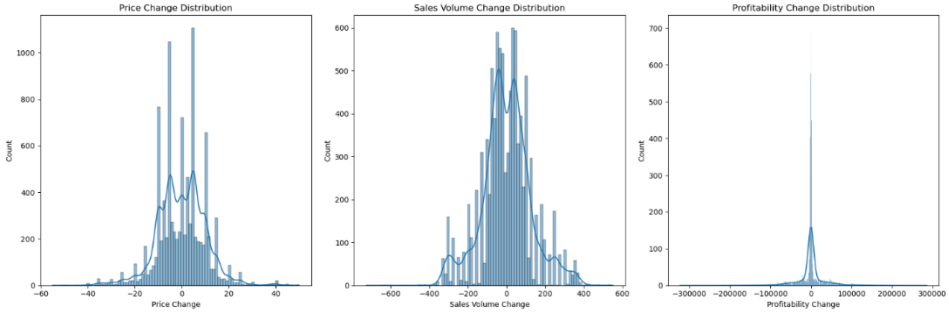
II. Monthly Sales Volume change trend
Sales volume sharply declines around October 2020, followed by fluctuating changes and an upward trend by late 2021, indicating increased sales activity.



III. Monthly Profitability change trend
Negative profitability spikes around October 2020 and January 2021, mirroring price and sales trends. Recovery follows, culminating in increased profitability by the end, indicating improved business performance.



7. Normal distribution of price change, sales volume change, and profitability change



Price changes cluster near zero with varied peaks, implying different strategies or seasonal adjustments.
Sales volumes vary around zero, mostly within -400 to 400 units, showing product or time-based variability.

8. Analysis of Variance (ANOVA)
Hypotheses Formulation
   I. Null Hypothesis (H0): There is no significant effect of seasonal patterns (such as month, quarter, day of the week) on sales trends.
   II. Alternative Hypothesis (H1): There is a significant effect of seasonal patterns on sales trends.

ANOVA Table for Invoice Date:

	sum_sq	df	F	PR(>F)
C(Invoice_Date)	2.887278e+11	723.0	2.814354	3.260110e-107
Residual	1.265858e+12	8921.0	NaN	NaN

ANOVA Table for Day of Week:

	sum_sq	df	F	PR(>F)
C(Day_of_Week)	1.163433e+10	6.0	12.112248	1.305106e-13
Residual	1.542952e+12	9638.0	NaN	NaN

	sum_sq	df	F	PR(>F)
C(Day)	2.535587e+10	30.0	5.313596	2.193640e-19
Residual	1.529230e+12	9614.0	NaN	NaN

ANOVA Table for Month:

	sum_sq	df	F	PR(>F)
C(Month)	3.572136e+10	11.0	20.595757	7.448268e-42
Residual	1.518865e+12	9633.0	NaN	NaN

ANOVA Table for Quarter:

	sum_sq	df	F	PR(>F)
C(Quarter)	1.876123e+10	3.0	39.257295	3.308267e-25
Residual	1.535825e+12	9641.0	NaN	NaN

Since PR(>F) (p-value) are all very close to 0, all five ANOVA tests show statistically significant results with very low p-values. This implies that the time periods have a significant impact on total sales of the products. Therefore, we can reject the null hypothesis in each case, suggesting that there are indeed differences in sales across different months, quarters, and days of the week.

6.0 Data Mining/Modelling

6.1 Data Modelling for Question 1

Linear regression and Random Forest Regression were chosen to answer the question 1.

Firstly, the 'Day of Week' and 'Quarter' features were encoded using one-hot encoding to convert them into numerical form suitable for regression analysis. Numerical features were scaled using StandardScaler() to ensure all features contribute equally to the model. The ColumnTransformer was used to apply these transformations and keep other columns intact.

6.1.1 Linear Regression

Linear regression was employed to understand the linear relationship between the transformed features and total sales.

Result:

```
Linear Regression RMSE: 12211.632002332222
Feature coefficients:
num__Year                    -2108.955828
num__Month                     197.474264
num__Day                      -248.046974
cat__Day of Week_Friday        996.783756
cat__Day of Week_Monday       -669.775137
cat__Day of Week_Saturday    -1089.326063
cat__Day of Week_Sunday      -1024.025048
cat__Day of Week_Thursday     2224.718382
cat__Day of Week_Tuesday      -168.218302
cat__Day of Week_Wednesday    -270.157589
cat__Quarter_2020Q1          -1361.304472
cat__Quarter_2020Q2           4662.650698
cat__Quarter_2020Q3           2434.433055
cat__Quarter_2020Q4          -5015.818205
cat__Quarter_2021Q1          -1952.173196
cat__Quarter_2021Q2           -491.306775
cat__Quarter_2021Q3           1594.779049
cat__Quarter_2021Q4            128.739846
dtype: float64
```

The linear regression model achieved an RMSE of approximately 12211.63, indicating the average difference between predicted and actual sales values.

Each coefficient represents the change in total sales for a one-unit change in the corresponding feature, holding other features constant.

- **Positive Coefficients**: Days like Friday and specific quarters such as 2020Q2 positively influence sales.
- **Negative Coefficients**: 2020Q1 and days like Sunday tend to have a negative impact on sales.

6.1.2 Random Forest Regression

A random forest regression model was also employed to capture non-linear relationships and interactions between features.

Result:

Random Forest Regression RMSE: 12739.197339241911

The random forest regression model yielded an RMSE of approximately 12739.2, slightly higher than the linear regression model. Random forest models provide feature important scores, which can identify which features have the most significant impact on sales and profitability.

6.2 Multiple Line Regression for Question 2

Independent Variables: 'Price Change' and 'Product'. The 'Product' variable was converted to dummy variables.
Dependent Variables include y_sales which is Sales Volume Change and y_profit which is Profitability Change

The dataset was split into training and testing sets with an 80-20 split. Two separate linear regression models were created using LinearRegression from sklearn, one model for predicting sales volume change, while another model for predicting profitability change.

We evaluated the models using the Mean Squared Error (MSE) and R-squared metrics:

- Mean Squared Error (MSE): Measures the average squared difference between the actual and predicted values. Lower values indicate better fit.
- R-squared: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher values indicate better fit.

Result:

```
Sales Volume Change Model
Mean Squared Error: 14610.981576769334
R-squared: 0.1315370564284809
Profitability Change Model
Mean Squared Error: 1578974173.66641
R-squared: 0.1485785049143058
```

For Sales Volume Change Model, the Mean Squared Error (MSE) is relatively high, indicating that the model's predictions are not very close to the actual values. The R-squared value of 0.1315 indicates that approximately 13.15% of the variance in sales volume change is explained by the model. This suggests a weak relationship between price changes and sales volume changes, indicating that other factors might be influencing sales volume more significantly.

For Profitability Change Model, the MSE is very high, indicating significant errors in the model's predictions for profitability changes. The R-squared value of 0.1486 indicates that around 14.86% of the variance in profitability change is explained by the model. This

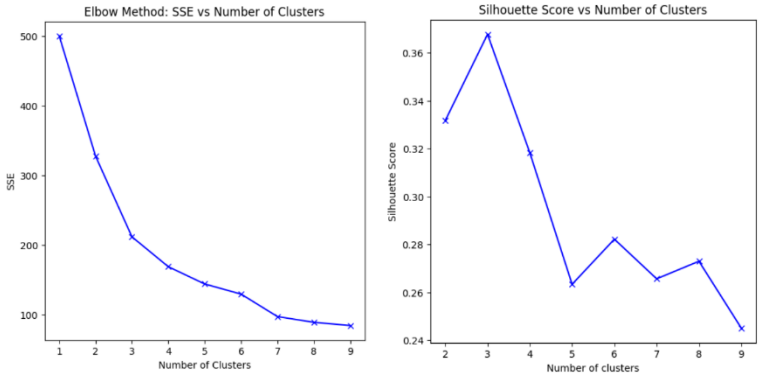
suggests a slightly stronger, but still weak, relationship between price changes and profitability changes compared to sales volume.

6.3 K-Means for Question 3

This is to find the K-Means clustering of price variation by states. We first identified the highest unit sold product. Then we generate the respective descriptive statistics and identify the average price per unit.

	State	mean_men_street	median_men_street	std_men_street	min_men_street	max_men_street	mean_women_apparel	median_women_apparel	std_women_a
0	Alabama	30.861111	29.5	13.963530	7.0	60.0	41.138889	38.0	14.3
1	Alaska	58.192308	61.0	12.289896	33.0	75.0	62.375000	67.0	10.5
2	Arizona	40.944444	41.0	7.270073	27.0	55.0	53.527778	54.0	11.7
3	Arkansas	34.277778	27.5	15.320595	15.0	70.0	40.000000	31.5	18.6
4	California	44.250000	44.5	9.499815	24.0	65.0	57.375000	59.5	12.7
5	Colorado	53.208333	55.0	11.923594	27.0	70.0	67.916667	73.5	15.7
6	Connecticut	45.694444	45.5	11.595121	23.0	65.0	48.777778	49.0	13.2
7	Delaware	49.416667	52.0	10.590877	32.0	65.0	53.666667	58.5	12.1
8	Florida	54.916667	55.0	13.721555	30.0	95.0	63.300000	63.5	18.0
9	Georgia	42.444444	42.0	7.658588	26.0	55.0	55.527778	57.5	12.8
10	Hawaii	58.583333	57.0	14.209356	36.0	85.0	71.625000	70.0	18.4

We determine the number of clusters to be 3 based on the results of elbow method and silhouette method. The data is scaled before performing K-Means clustering.

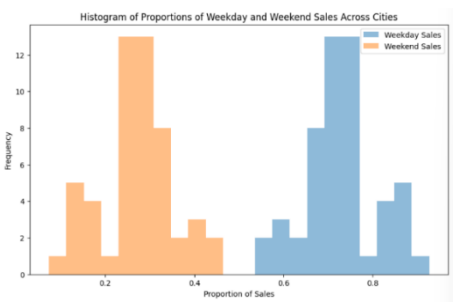


6.4 Hypothesis Testing for Question 4

To identify the significant difference in total sales across cities between weekdays and weekends, we first group the data based on cities and split it into two parts: weekday sales and weekend sales. The proportion column is added for histogram plotting. From the histogram, it is shown that weekday sales are skewed towards higher proportions (0.6 to 0.8), while weekend sales are skewed towards lower proportions (0.2 to 0.4). Both weekday and weekend sales are not normally distributed.

	City	Weekday_Weekend	Total Sales	Proportion
0	Albany	Weekday	2485660.0	0.830129
1	Albany	Weekend	508644.0	0.169871
2	Albuquerque	Weekday	1843947.0	0.652808
3	Albuquerque	Weekend	980694.0	0.347192
4	Anchorage	Weekday	1211189.0	0.669007
...	...	...	...	...
99	St. Louis	Weekend	286579.0	0.240921
100	Wichita	Weekday	823376.0	0.671971
101	Wichita	Weekend	401938.0	0.328029
102	Wilmington	Weekday	1050610.0	0.696443
103	Wilmington	Weekend	457927.0	0.303557

104 rows × 4 columns



The Levene test results also indicate that weekday and weekend sales have unequal variances.

```
Levene test results:
Statistic: 32.8655969529778, p-value: 1.0108273247406994e-07
The variances are significantly different from each other
```

Hence, we use the Mann-Whitney U test to identify the difference in total sales at 0.05 significant level.

7.0 Results and Product Discussions

7.1 Results and Discussions for Question 1

To answer the question "How do seasonal trends affect sales and profitability across different product categories?", several observations are discussed as follows:

Observations from Seasonal Sales Patterns:

1. Yearly Sales Trends:
    - Sales in 2021 were significantly higher than in 2020, with peak sales occurring around August 2021.
2. Product Trends:
    - Street footwear consistently showed higher sales compared to other categories.
    - Athletic footwear had lower sales relative to other product categories.
3. Day of the Week Trends:
    - Fridays had the highest sales, followed by Tuesdays.
    - Mondays had the lowest sales.
4. Day of the Month Trends:
    - Certain days of the month, specifically the 17th, 10th, and 24th, had higher sales.
5. Monthly and Quarterly Trends:
    - July had the highest sales each year.
    - The third quarter consistently showed the highest sales.

Observations from Correlation Analysis:

1. High Correlation with Total Sales:
    - Units sold, operating profit, and price per unit were highly correlated with total sales.
2. Sales Volume and Profitability:
    - A strong positive correlation was observed between sales volume changes and profitability changes.

Observation from Monthly Trends:

- Price Changes with notable fluctuations were observed, with significant declines around January 2021 followed by recoveries, indicating regular price adjustments.
- Sales Volume Changes has sharp declines early in the observed period (October 2020) were followed by fluctuating periods of increases and decreases.

- Profitability Changes with significant negative changes were noted around October 2020 and January 2021, with a trend of recovery towards the end of the period.

Observation from Hypothesis Testing (ANOVA):

- Significant seasonal patterns were found across different months, quarters, and days of the week.
- ANOVA results with very low p-values confirmed that seasonal trends significantly impact sales, allowing rejection of the null hypothesis.

Observation and Results from Data Modeling:

1. Linear Regression:
    - Achieved an RMSE of approximately 12211.63
    - Analysis of feature coefficients revealed positive impacts from Fridays and certain quarters, and negative impacts from Sundays and certain years.
2. Random Forest Regression:
    - Achieved an RMSE of approximately 12739.2, slightly higher than linear regression.
    - Provided additional insights into feature importance.

Seasonal trends significantly influence sales and profitability, with notable variations observed across different days, months, and quarters. The exploratory data analysis and hypothesis testing revealed clear seasonal patterns, while data modeling further confirmed the substantial impact of these patterns on sales. Both linear regression and random forest regression models highlighted the importance of understanding seasonal effects to predict sales more accurately.

7.2 Results and Discussions for Question 2

To answer the question “How does the pricing of products influence both sales volume and profitability across different types of products?”, several observations are discussed as follows:

Observation of Correlation Analysis from EDA:

- Units Sold, Total Sales, and Operating Profit is highly correlated with each other.
- Price per Unit positively correlated with total sales and operating profit.
- Sales Volume and Profitability has strong correlation of 0.73, indicating a significant relationship between these two variables.

Observation from Monthly Trends:

- Price Change has notable decline in January 2021, followed by recovery with regular fluctuations.
- Sales Volume Change has sharp decline in October quse2020, followed by fluctuating increases and decreases, ending with an upward trend in late 2021.
- Profitability Change has significant negative changes around October 2020 and January 2021, with an upward trend towards the end of the period.

Observation from Normal Distribution Analysis:

- Price Changes are centered around zero with multiple peaks.
- Sales Volume Changes are lightly skewed towards negative changes with multiple modes.
- Profitability Changes sharply peaked around zero with long tails indicating potential significant changes.

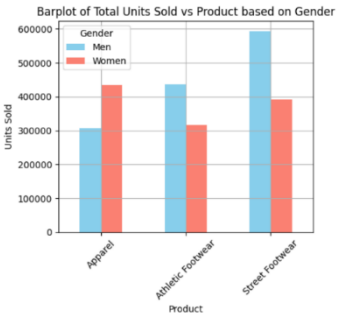
Observation of Data Modeling through Multiple Linear Regression:

- Sales Volume Change Model with R-squared around 0.1315 indicates that approximately 13.15% of the variance in sales volume change is explained by the model. This suggests a weak relationship between price changes and sales volume changes, indicating that other factors might be influencing sales volume more significantly.
- Profitability Change Model with R-squared around 0.1486 indicates that around 14.86% of the variance in profitability change is explained by the model. This suggests a slightly stronger, but still weak, relationship between price changes and profitability changes compared to sales volume.

The multiple linear regression analysis shows that price changes have a relatively small impact on both sales volume and profitability changes across different product types. The low R-squared values suggest that there are other significant factors affecting these outcomes that are not captured in the current model.

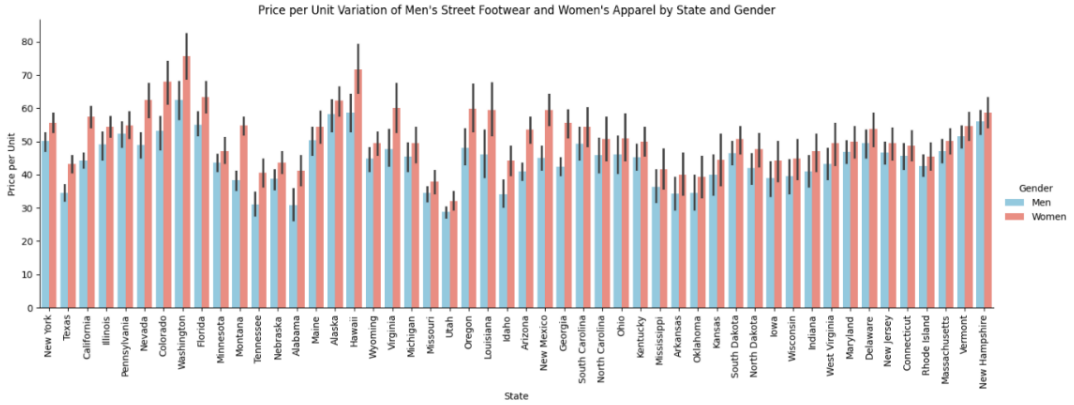
The EDA indicated several trends and correlations, but the regression models highlight the complexity of predicting sales volume and profitability changes based solely on price changes and product type.

7.3 Results and Discussions for Question 3



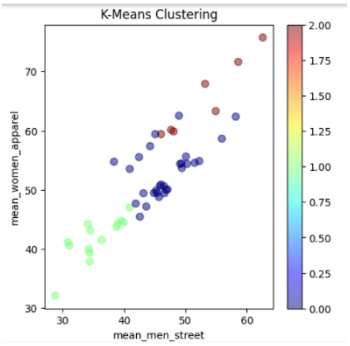
From the bar plot of Total Units Sold vs Product based on Gender, we can identify the highest total units sold product category for Men is Street Footwear and for Women, it is Apparel.

Based on the bar plot of Price per Unit Variation of Men's Street Footwear and Women's Apparel by State and Gender referenced from (Waskom, 2021), it is shown that in most states, the price per unit for Women's Apparel is higher than Men's Street Footwear.



States such as Washington, Hawaii and New Hampshire show higher prices per unit for both product categories while Missouri, Utah and Tennessee have relatively lower prices per unit.

The scatter plot of K-means Clustering, average price per unit of Men's Streetwear and Women's Apparel are divided into three clusters (0, 1 and 2). Cluster 0 contains states with moderate average price per unit for both Men's Streetwear and Women's Apparel, Cluster 1 includes states with lower average prices for both product categories, and Cluster 2 includes states with higher average prices. States within the same cluster exhibit similar price trends for men's footwear and women's apparel.

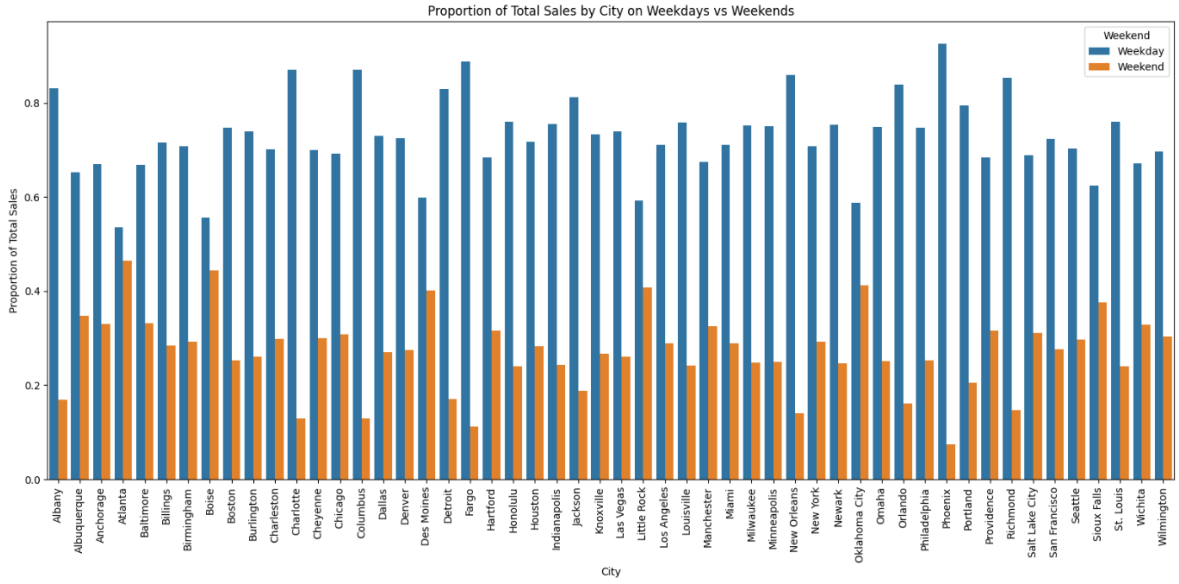


7.4 Results and Discussions for Question 4

From the Mann-Whitney U test to test for the significant difference of total sales, it is shown that p-value is lesser than alpha = 0.05. Hence, there is a significant difference in total sales between weekdays and weekends.

```
Mann-Whitney U test results:
Statistic: 2433.0, p-value: 1.0108273247406994e-07
There is a significant difference in total sales between weekdays and weekends.
```

The significant difference in total sales between weekdays and weekends in cities is also shown in the bar plot below. A larger proportion of total sales occurs on weekdays compared to weekends. This trend is consistent but varies in magnitude from city to city.



8.0 Challenges and Limitations

Challenges for question 1 include data inconsistencies affecting seasonal patterns, with unaccounted factors like market conditions and external events potentially distorting trends. Limitations involve models like linear regression and random forest showing limited predictive power. Future analysis could use ensemble methods and include additional variables to better capture and predict seasonal fluctuations accurately.

Challenges in Question 2 stem from the complexity of interdependencies among price changes, sales volume, and profitability, revealing that the current model may inadequately capture these relationships due to omitted factors. Limitations include low R-squared values in regression models, signifying unexplained variance in sales and profitability, likely due to insufficient attributes such as promotional activities not being considered in the analysis.

In Question 3, regarding price per unit variation across states, if there is a need to implement more dimensions in the future, K-means might not be a suitable approach for this due to its scalability and sensitivity to outliers (Pandey, 2024).

In the hypothesis testing of weekday and weekend sales in Question 4, the reliability and accuracy of the data source can be a limitation. Any errors in data collection or recording could impact the validity of the testing results.

9.0 Conclusion

This analysis project of the Adidas US Sales Dataset from Kaggle (Adidas Sales Dataset, 2022) aimed to enhance operational efficiency and business agility for Adidas in the US market. By addressing the research questions, we gained valuable insights into the various factors influencing sales and profitability.

The analysis identified strong seasonal patterns influencing sales and profitability across regions and products. Sales peaked in August 2021, with Fridays consistently showing higher sales and Mondays lower. While correlations between total sales, units sold, and operating profit were significant, predictive models indicated seasonal trends alone are insufficient for accurate forecasts. Integrating additional variables is crucial for improving predictive accuracy.

Multiple linear regression showed minimal impact of price changes on sales volume and profitability, suggesting other factors are significant. Exploratory analysis highlighted trends but predicting sales and profitability solely from price and product type is challenging. Incorporating additional variables is essential for a more comprehensive understanding.

Furthermore, we identified significant variations in the highest total units sold across states based on distinct gender preferences and price-per-unit differences within these categories across various states. These findings underscore Adidas's need to adopt localized pricing strategies tailored to regional market dynamics (Li, 2023, p. 35).

Simultaneously, exploring sales trends across cities over time highlighted a notable difference in weekday and weekend sales, emphasizing the potential for Adidas to implement tailored promotional activities for each city (Li, 2023, p. 36). Aligning marketing efforts and inventory levels with consumer behavior patterns can maximize sales opportunities and enhance operational efficiency across its retail locations in the US.

10.0 References

Adidas Sales Dataset. (2022, December 23). Kaggle https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6(60)*, 3021. https://doi.org/10.21105/joss.03021

Pandey, A. K. (2024, February 28). *A Simple Explanation of K-Means Clustering*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/

Li, X. (2023). Analysis of the Marketing Strategy of Adidas. *Advances in Economics, Management and Political Sciences*, *23*(1), 33–38. https://doi.org/10.54254/2754-1169/23/20230346