Data Analysis Portfolio

I am from “Sunny Technologies”. I am a resident of India. I have completed my “Master’s in Electrical Engineering” from “University of North Texas”. I had worked as an “Associate Engineer” at Samsung in IOWA, USA and as “Field Test Engineer” at P3 Communications in Texas, USA. Presently, I am pursuing Data Analyst career through online courses. My goal is to become a Data Scientist and explore Machine Learning and Artificial Intelligence world. I have developed skills in Python, SQL, Power BI, Tableau, Excel. I consider myself as an Intermediate level user in them. My hobbies include playing badminton, squash.

Professional Background

SUNNY TECHNOLOGIES

DATA ANALYST

ABOUT ME

I am a working professional seeking a job as a Data Analyst. I want to work in a challenging environment where I can put all my capabilities to use, reach my potential and expand my knowledge, both professionally and personally

EDUCATION

University of North Texas
2017 - 2018
[Master's in Electrical Engineering.]

Chaitanya Bharathi Institute of Technology
2011 - 2015
Bachelor's in Electrical and Electronic Engineering

WORK EXPERIENCE

Samsung Associate Engineer
Jun 2019–Feb 2020
Working on mobile handsets with US Cellular Carrier. Maintaining quality in mobiles, reporting bugs, training new hires.

P3 Communications Field Test Engineer
Mar 2018–May 2019
Worked on filed testing carrier signal capabilities.

SKILLS

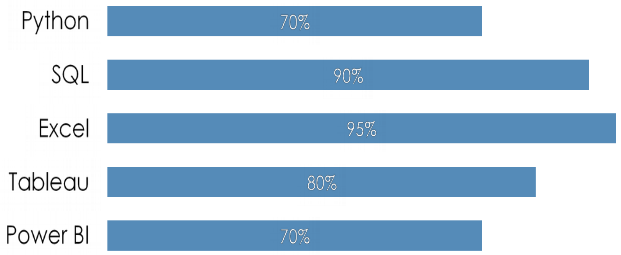


Table of Contents

Udemy Project Description

Situation : Working as a Data Analyst at an educational tech company Udemy, I had to complete a challenge provided by my manager within a deadline of three weeks. I was provided with historical data containing some of the courses offered by the company, how much income had they generated, number of subscribers and rating of the course.

Task : The challenge was to perform analysis on the data, to check why the revenue is falling and to provide valuable insights to improve the revenue of the company for the next quarter.

Action : Data provided to us was raw data. So, to make it useful for data analysis, I had to perform some necessary data cleaning with the help of '*Excel'* tool. Data Cleaning included removing blank rows, removing duplicate values. Also, I have created columns to show if the course is a 'Free course' or 'Paid course'. The four courses(web development, Graphic design, Musical Instruments and Business Finance) were in different sheets, so merged for a clean data organization. Used *Tableau*, a visualization tool to provide some more visual insights by creating dashboards and stories. Lastly, to find the root cause of the problem, I have used '*five why's analyses*.

Results: From the analysis and visualizations, it is evident that revenue is failing due to decrease in the number of subscribers. The subscribers are leaving because the rating of the courses was also declining over the years. Insights were provided to add more courses from reputed tutors and to find what all courses are in demand at present.

The Problem

Business Problem : The problem that we are trying to solve is finding out what topics or courses are in demand and the necessary steps to be taken to increase the revenue of the company in the next quarter.

Time to finish the project : It took 14 days to complete the project and a buffer period of 3 days if any queries arise.

Timeline for Data Analysis Project

Process	**Time**
Business Understanding	2 days
Data Collection	2 days
Data Preprocessing	2 days
Data Analysis	3 days
Data visualization and presentation	3 days
Documentation	2 days
Buffer Period (For feedbacks and necessary changes in the project)	3 days
Total	17 days

What data should be collected to understand this problem? How should it be presented?

The data used for this project is provided by *'Entry Level'* Organization as an excel format. They provided us four sheets containing data on Design, Business, Music and Web development courses. Each Sheet contains course_id, course_title, URL, price, num_subscribers,num_reviews,num_lectures,level,Rating,content_duration,published_timestamp ,subject columns. Combining all data sheets into one spreadsheet; it contains 58,832 rows of data.

Using Tableau and Pivot Tables from Excel, data is presented by creating dashboards and stories.

What questions would you ask to better understand the business problem?

1. By what percentage do you want to increase the revenue.
2. How do you want to increase the revenue.(For Ex: Increase price of course in demand, to attract more subscribers add discount on course, adding more courses of the in-demand category taught by highly reputed organization or tutor)?
3. Do you want to know what courses are in demand at present?

4. Why do you want to track the performance of each course?
5. Do you want to know which region has highest enrollment for each course or category?

Design

Data Cleaning:

1. The data doesn’t have any missing values. This is achieved by conditional formatting. Going to conditional formatting and selecting highlight cells rules, click on equal to and enter null. This will highlight in red if there are any null values.
2. There are some duplicate values in the data. All duplicate values are removed. Select course_id column and on ribbon go to ***data*** tab. In the data tools, click on remove duplicates. This will remove all the duplicates in the column.
3. In Web Development sheet, the subject column contained *subject* as the prefix. The prefix is removed by selecting the column and using **‘left’** formula. And can also be achieved by using **‘find and replace’** option.

Data Organization

1. Since published timestamp column has both date and time, a separate date column is created. Date column is separated using “**left**” formula. “**LEFT(row_num, number of letters)**”.
2. Free course column is created to find if the given course is free or paid course. Formula used for this is **“=IF(D2=0,"Free Course", "Paid Course")”.**
3. Free beginner course column is created to find if the course is a beginner level free course or not. Formula used for this is **“=IF(AND(D2=0,H2="Beginner Level"),"Free Beginner Course", "Not Free Beginner Course")”.**
4. All courses free or paid column is created to find the course level along with if the course is paid or free. Formula used for this is “**=IFS(AND(D2=0,H2="Beginner Level"),"Free Beginner Course", AND(D2=0,H2="All Levels"),"Free All Level Course", AND(D2=0,H2="Intermediate Level"),"Free Intermediate Level Course", AND(D2=0,H2="Expert Level"),"Free Expert Level Course", AND(D2<>0,H2="Beginner Level"),"Paid Beginner Course", AND(D2<>0,H2="All Levels"),"Paid All Levels Course", AND(D2<>0,H2="Intermediate Level"),"Paid Intermediate**

Level Course", AND(D2<>0,H2="Expert Level"),"Paid Expert Level Course")" .

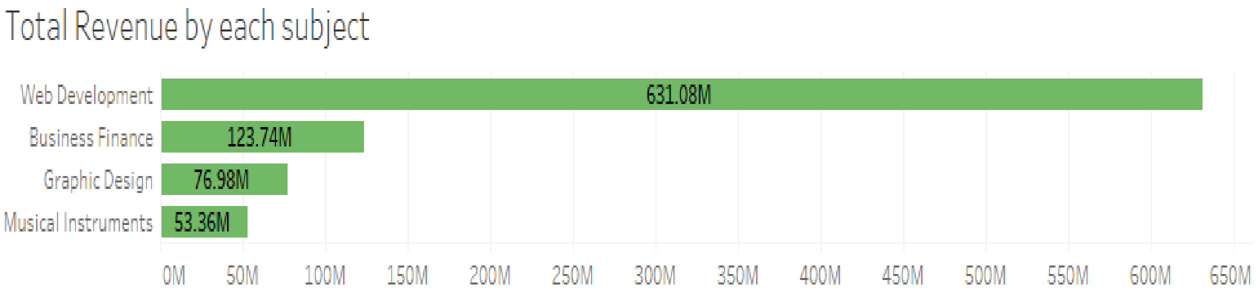
5. All the spreadsheets are merged into one single spreadsheet so that it will be easy to work on pivot tables.

Findings

Revenue

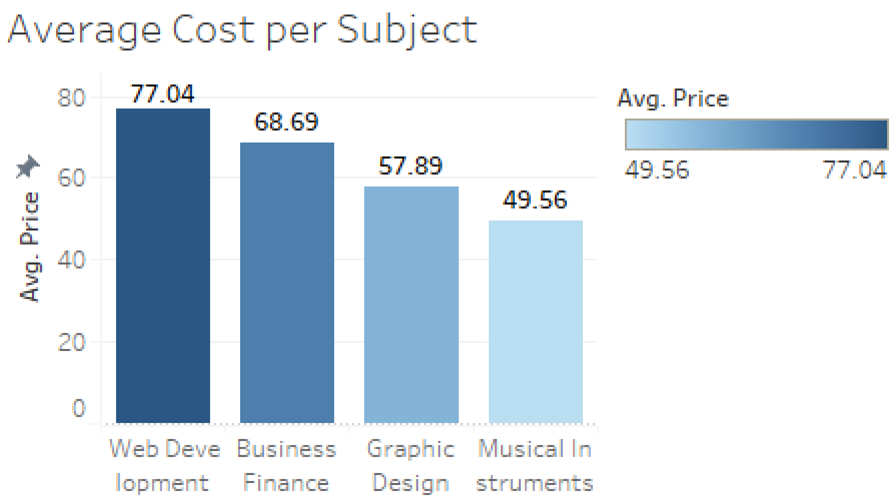
1. Overall, total revenue generated by the company from the given data is **885,160,005(885.16million).**

2. Out of all the subjects, Web development courses contributed for about **71%** of the revenue.

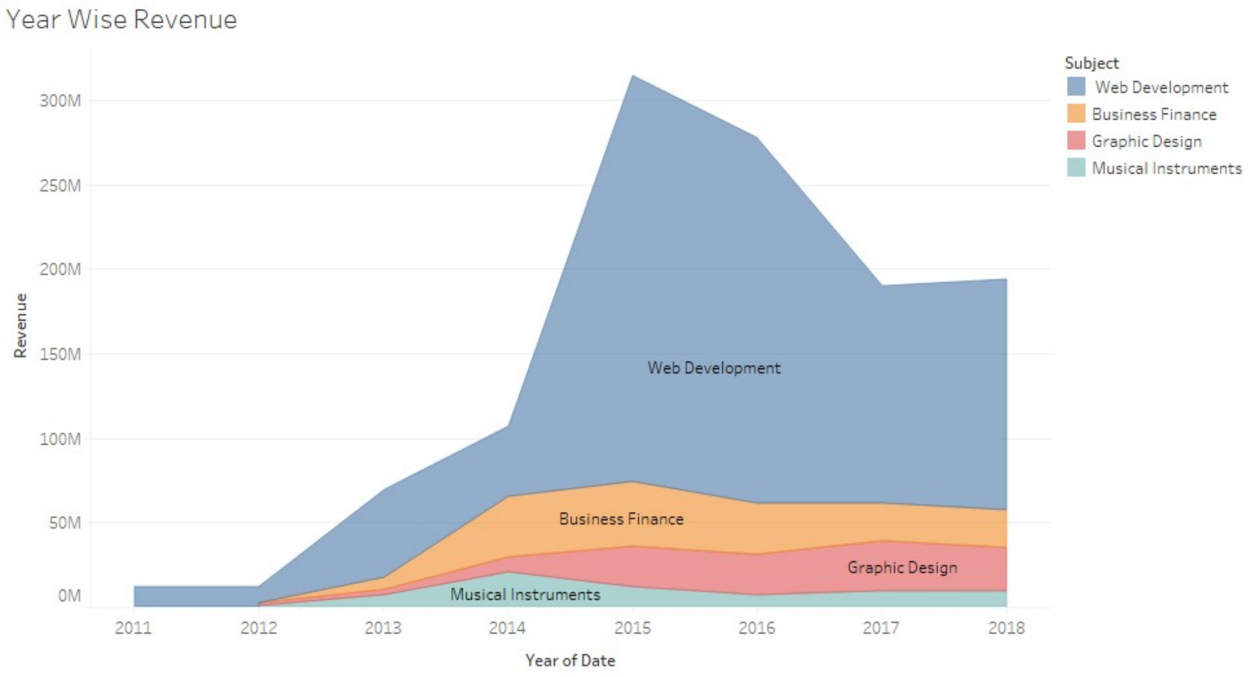


Sum of Revenue for each Subject. The marks are labeled by sum of Revenue. The data is filtered on Action (Subject) and Date Year. The Action (Subject) filter keeps 4 members. The Date Year filter keeps multiple members.

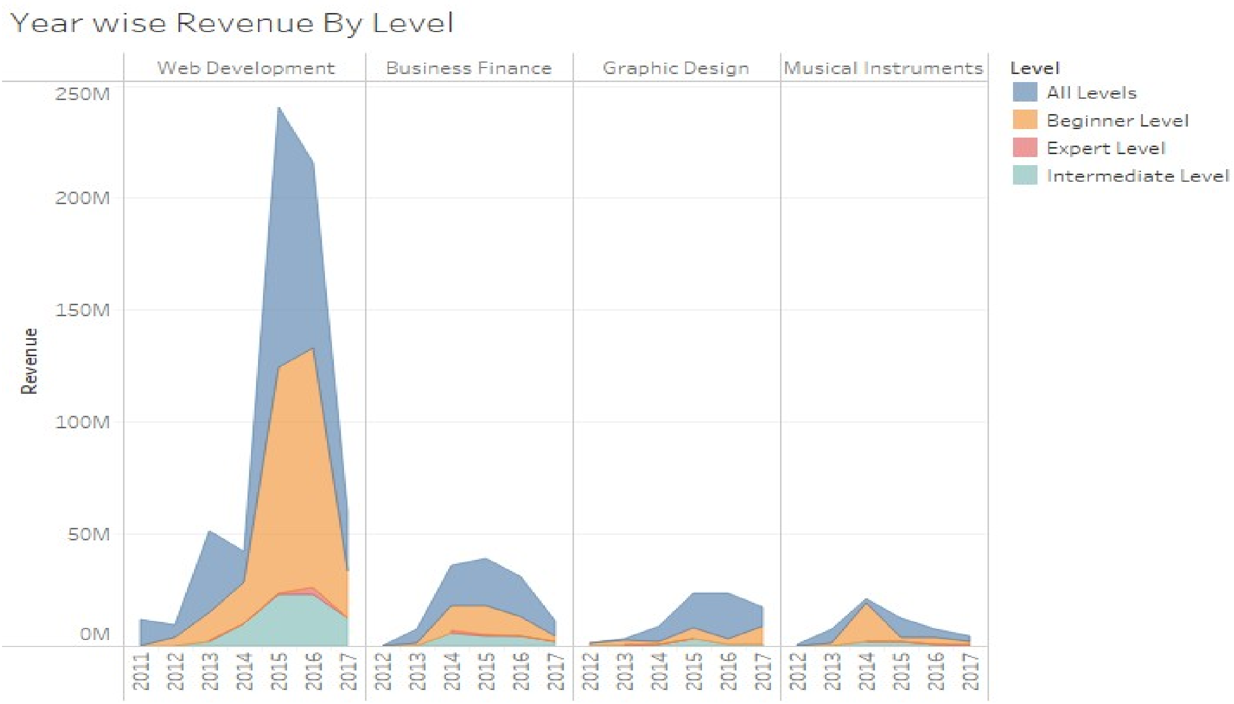
3. Also, web development courses have the highest average cost per subject.



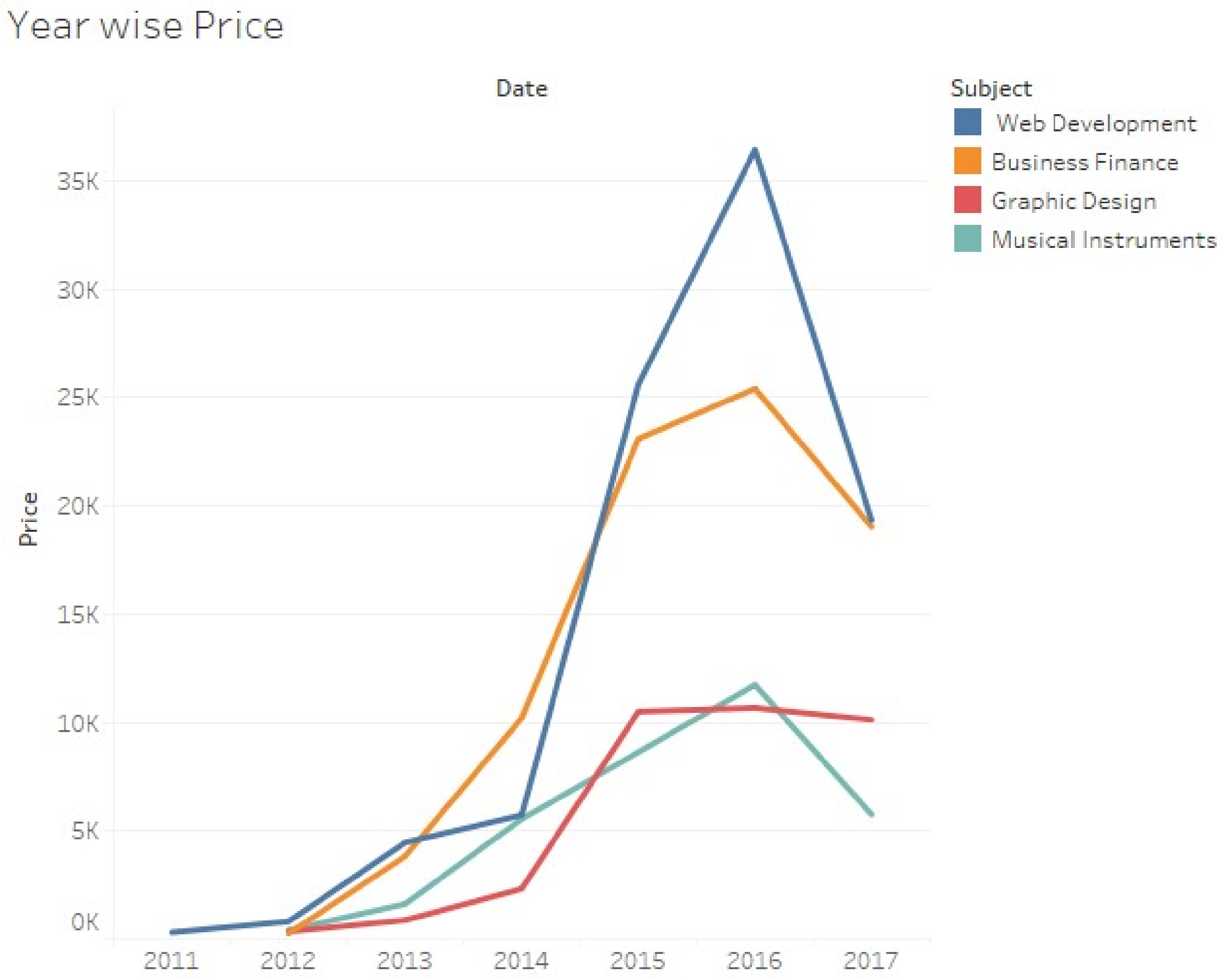
4. By plotting year wise revenue for each subject, revenue is decreasing for all the given subjects. All the subject's revenue peaked in the year 2015 and then consistently fell.



5. By plotting year wise revenue by each level of course, all level courses and beginner level courses account for most of the revenue.

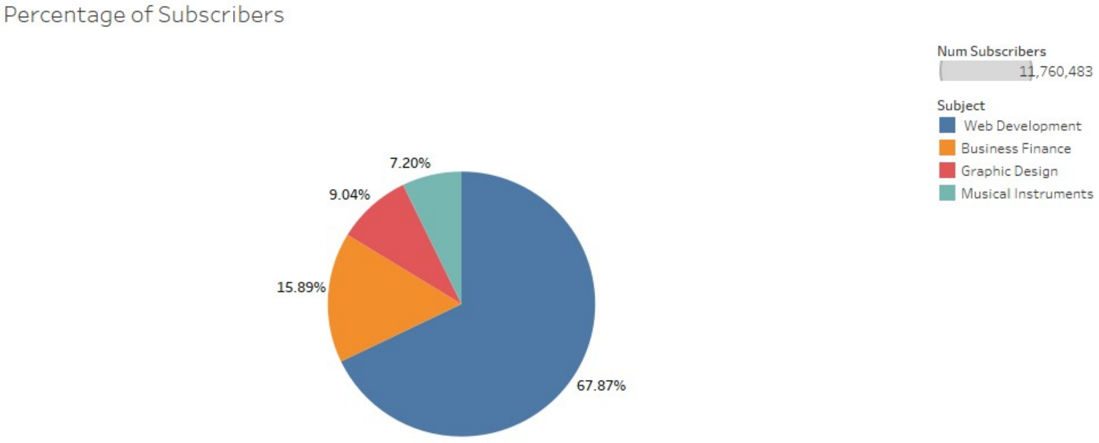


6. From year wise price chart, we can see that price of the courses have been decreased. There are more subscribers even when the price is high.

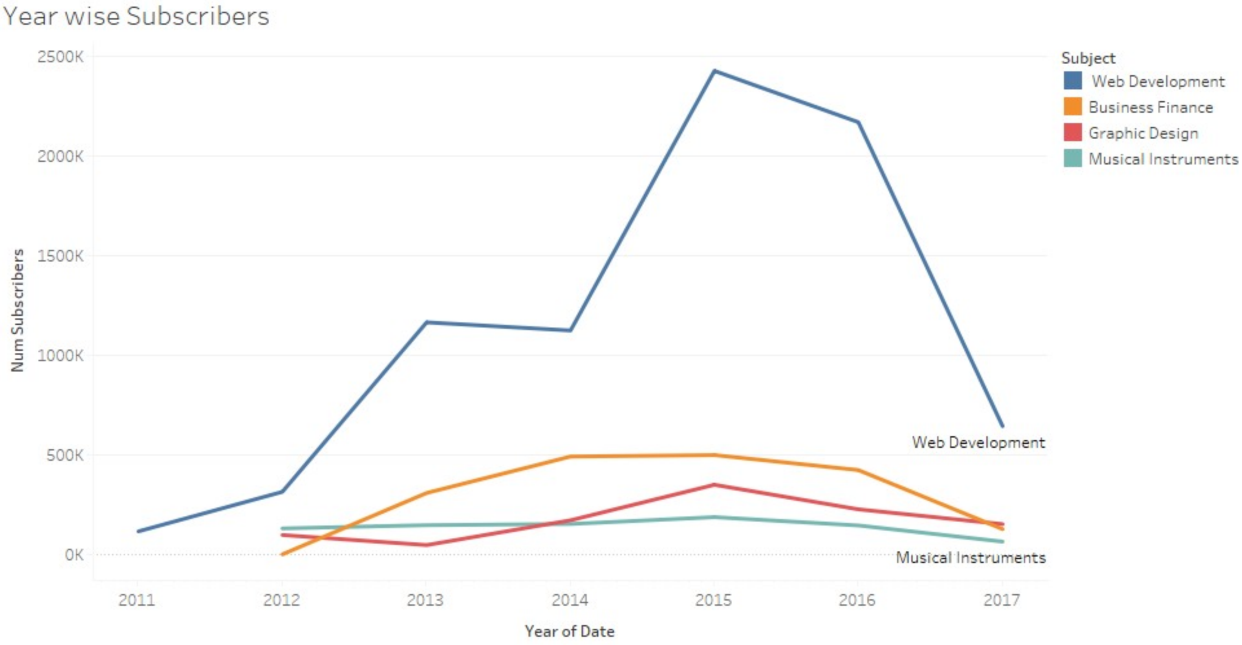


Subscribers

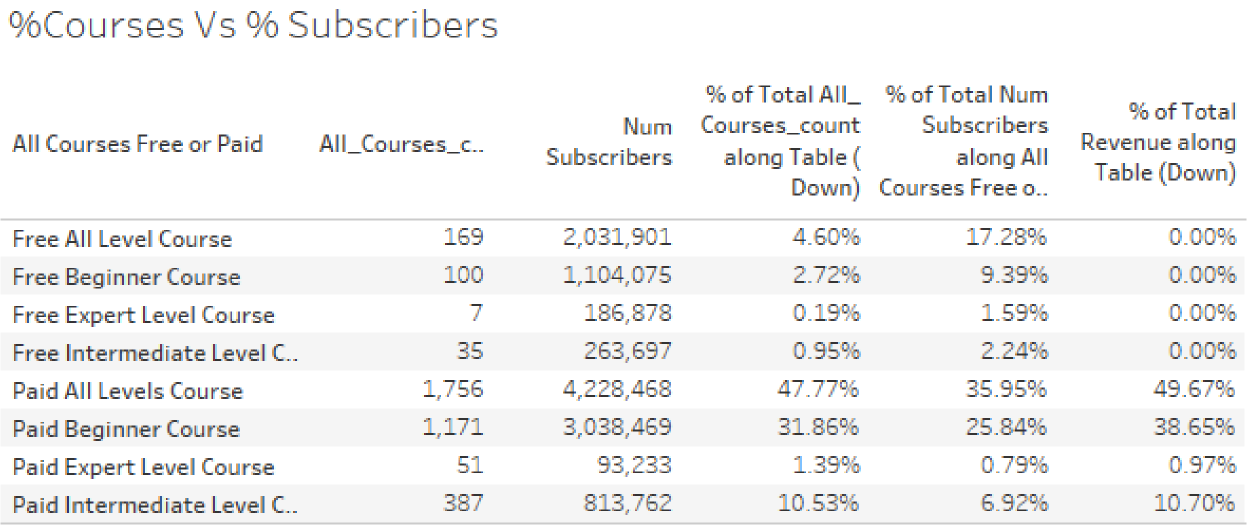
1. There about **11,760,483** Subscribers for all the courses.
2. Web development subscribers account for a total of **67.87%** of the total subscribers making it the highest. About **7.2%** for musical instruments courses making, it the lowest.


Percentage of Subscribers

3. By plotting year wise subscribers, number of subscribers are constantly decreasing. In the year 2015, the subscribers count was more than all other years for all the subjects. There is a drastic drop in web development course subscribers from a peak of about 2.5M subscribers to 500K.


Year wise Subscribers

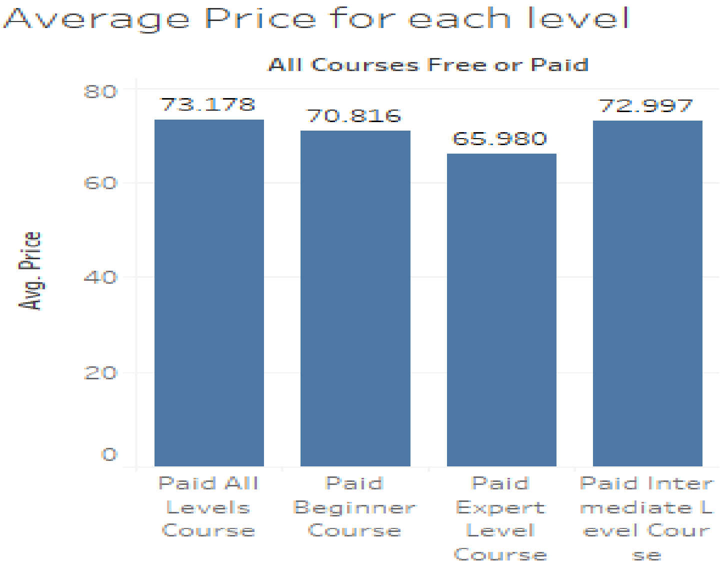
4. From the below table, we can deduce that the majority percentage of subscribers preferred paid all level and paid beginner level courses. Also, the total number of courses for these levels is also high. Although, there are only about **10.5%** of paid intermediate level course, it accounts for about **10.7%** of the revenue.



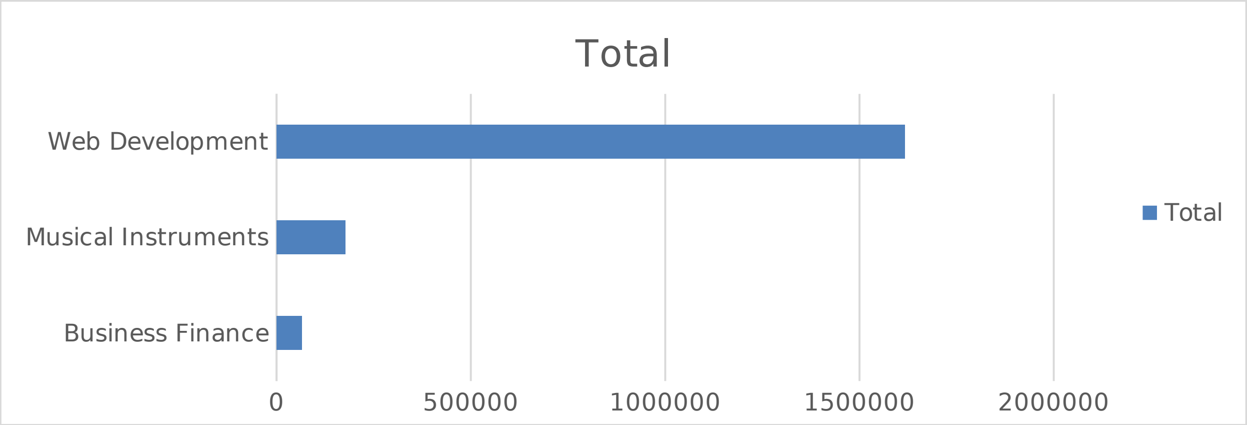
%Courses Vs % Subscribers

All Courses Free or Paid	All_Courses_c..	Num Subscribers	% of Total All_ Courses_count along Table (Down)	% of Total Num Subscribers along All Courses Free o..	% of Total Revenue along Table (Down)
Free All Level Course	169	2,031,901	4.60%	17.28%	0.00%
Free Beginner Course	100	1,104,075	2.72%	9.39%	0.00%
Free Expert Level Course	7	186,878	0.19%	1.59%	0.00%
Free Intermediate Level C..	35	263,697	0.95%	2.24%	0.00%
Paid All Levels Course	1,756	4,228,468	47.77%	35.95%	49.67%
Paid Beginner Course	1,171	3,038,469	31.86%	25.84%	38.65%
Paid Expert Level Course	51	93,233	1.39%	0.79%	0.97%
Paid Intermediate Level C..	387	813,762	10.53%	6.92%	10.70%

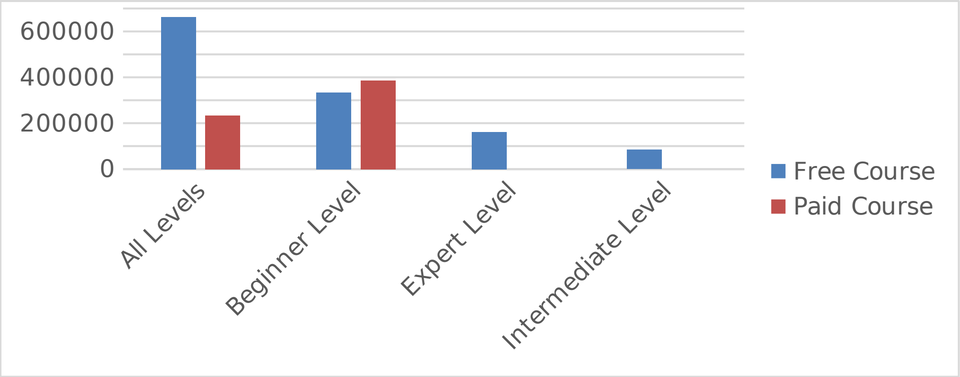
5. Considering Average price for each level chart, paid all level and paid intermediate courses have the highest average cost. This implies that number of subscribers are not dependent on price.



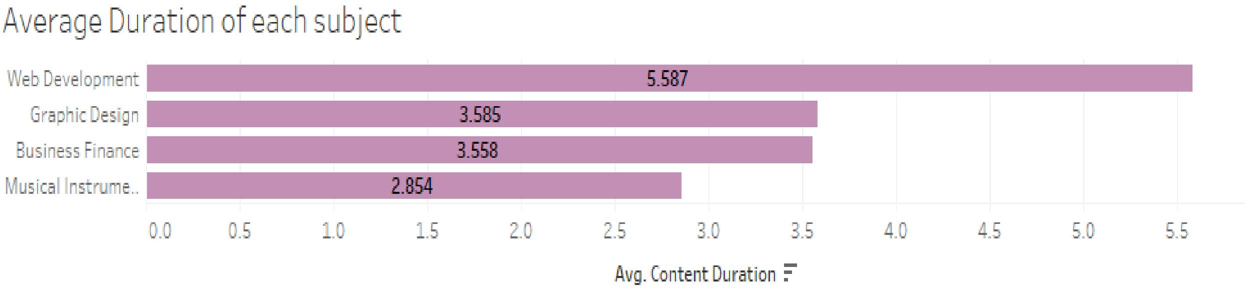
6. In top 20 most subscribed courses, web development courses are more.



7. Free courses are most subscribed among the top 20 courses. But Paid beginner level courses are more subscribed than free beginner courses.



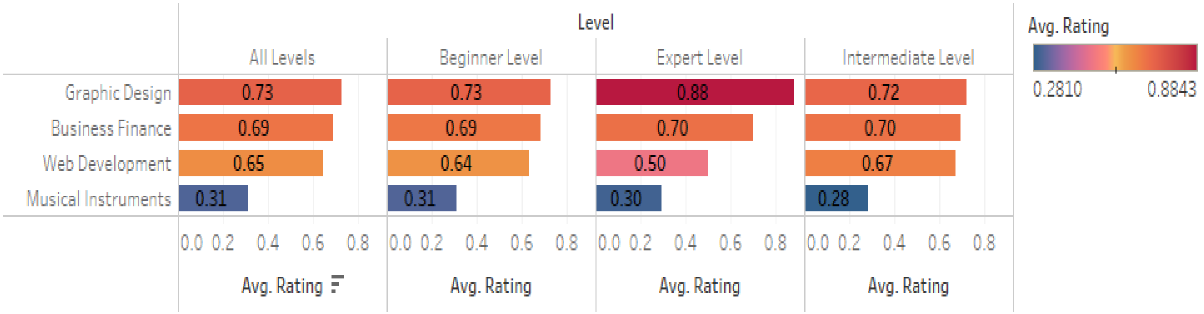
8. Average Duration for each subject, web development has more content duration.



Ratings

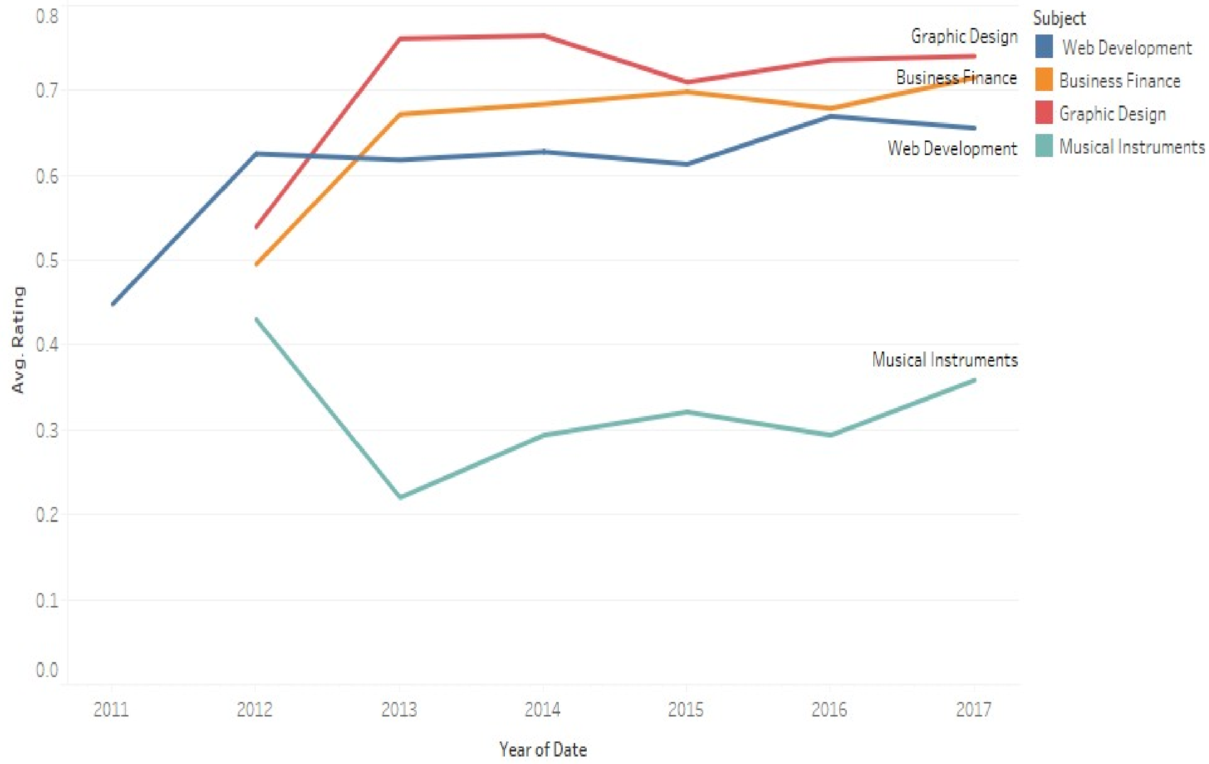
1. The average rating per each subject is less than 1. There is a serious need to improve the average rating. Musical instruments courses have the worst rating of about 0.3 for each level.

Average rating for each subject by Level



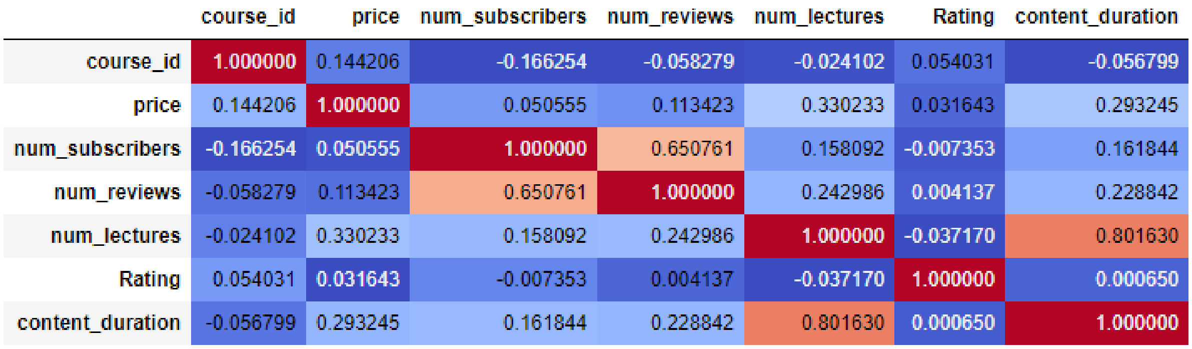
2. From year wise rating chart, rating for web development courses has been on constant decrease while graphic and business courses improved.

Year wise Rating



Analysis

1. From the correlation chart, we can find that there is no strong relation between price and number of subscribers. So, price is not a deciding factor.
2. There is strong relation between number of subscribers and number of reviews.



	course_id	price	num_subscribers	num_reviews	num_lectures	Rating	content_duration
course_id	1.000000	0.144206	-0.166254	-0.058279	-0.024102	0.054031	-0.056799
price	0.144206	1.000000	0.050555	0.113423	0.330233	0.031643	0.293245
num_subscribers	-0.166254	0.050555	1.000000	0.650761	0.158092	-0.007353	0.161844
num_reviews	-0.058279	0.113423	0.650761	1.000000	0.242986	0.004137	0.228842
num_lectures	-0.024102	0.330233	0.158092	0.242986	1.000000	-0.037170	0.801630
Rating	0.054031	0.031643	-0.007353	0.004137	-0.037170	1.000000	0.000650
content_duration	-0.056799	0.293245	0.161844	0.228842	0.801630	0.000650	1.000000

Five Why's Analysis

Quoting the business problem, there is a fall in revenue over the years and the company want to improve its revenue in the next quarter.

Hypothesis: The rating of the courses is falling hence the subscribers are leaving.

1. Why there is a peak revenue for the web development courses?

A: The number of courses is more compared to other courses and the demand for web development is high.

2. Why is there a decrease in revenue?

A: Because there is a decrease in number of subscribers. Also, there is a drop in price.

3. Why there is a decrease in subscribers?

A: The number of courses is less compared to previous years.

4. Why is there a decrease in number of courses?

A: The rating of the courses is constantly decreasing.

5. Why there is a drop in rating?

A: Students did not like the course content.

Insights

- To attract more subscribers, the rating of the courses needs to be improved a lot. This can be achieved by collaborating with highly reputed instructors and making the course content in par with the real-world standard.
- Add more paid intermediate level courses. This will in turn generate revenue.
- Add more courses overall levels.
- Increase the price for all levels of courses with more rating.

Conclusion

The data analysis for the Udemy dataset is successfully completed. All the data given is cleaned, sorted, filtered, and organized for analysis. From the analysis, it can be concluded that rating of all the courses should be improved and price of the courses with higher rating can be increased.

Walmart Retail Analysis Project Description

Situation : Working as a data analyst at Sunny technologies, I was presented with a challenge to work on Walmart historical data containing about data from 45 of their stores and provide data analysis. Walmart is facing issues on satisfying demands of customers as they run out of stock sometimes during certain holidays and events. They want to maintain a steady supply Vs demand so there won't be any loss in retail and keep the customer satisfaction high.

Task: The task is to provide exploratory data analysis on Walmart data and present with insights on how the business effects with certain events and holidays so they will be prepared in the future to stock the stores with appropriate supply.

Action : The data given to me was in raw form in a .csv format. First the data is cleaned, sorted, and is made ready for analysis. Several data questions were asked to further understand the business problem. Pivot tables are created to understand relation between sales and several other factors. Appropriate charts were created to represent the data visually. Dashboards and stories are created in Tableau which can be used for presentation.

Results: From the analysis, I have found that the sales are most in the month of December and 2011 had seen greater sales. There is no definite correlation between sales and the factors that were provided like temperature, fuel price etc. The standard deviation of weekly sales across all years is 564366.622 which implies that there is a lot of difference in the revenue generated among the stores. The average sales for Thanksgiving holiday are greater than any other holidays or events and non-holiday sales. Store 20 performed better than all other stores while store 33 recorded least sales.

The Problem

Business Problem : The Walmart organization is losing their business and customers as they couldn't meet demand on certain holidays and events. The objective is to determine the factors affecting the sales and to analyze the impact of markdowns around holidays on the sales.

Time taken to finish the project: It took me 7 days to finish the project.

Timeline for Data Analysis Project

Process	**Time**
Business Understanding & Data Collection	1 day
Data Preprocessing	1 day
Data Analysis	2 days
Data visualization and presentation	2 days
Documentation	1 days
Total	7 days

What data should be collected to understand this problem? How should it be presented?

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in which you will find the following fields:

- Store - the store number
- Date - the week of sales
- Weekly_Sales - sales for the given store
- Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
- Temperature - Temperature on the day of sale
- Fuel_Price - Cost of fuel in the region
- CPI – Prevailing consumer price index
- Unemployment - Prevailing unemployment rate

What questions would you ask to better understand the business problem?

1. What policies have been implemented in the past and why it has failed?
2. What are the most sold items in the store?
3. How satisfied is the customer with the store and staff?
4. What are the discounts and promotions used during holidays and events?
5. Who is our customer and where do they live?

Design

Data Cleaning

1. The date format provided is in wrong format. To make it in right format, selected the date column and clicked on *‘data’* tab on the ribbon and then selected *‘text to columns’* option. Select delimited option and click next. Unselect all the delimiter options and click next. In the column data format, click on date and from dropdown menu select ‘DMY’ format and click finish. Now all the dates are formatted as “DD/MM/YYYY’.
2. Changed the data types of all numerical columns to number format as they are in general format.
3. Created a column that specifies which event or holiday on the specific date. Used ‘MATCH’ formula to search for dates and it will return an index of the row number where the date is located. Using the row number, a new entry is created with the specific holiday or event name.

Visualization tools

1. Used Excel to create pivot tables and some charts. Excel is a spreadsheet software, and its biggest benefits is the ability to organize large amounts of data into orderly, logical spreadsheets and charts. With the data organized, it's a lot easier to analyze and digest, especially when used to create graphs and other visual data representations.
2. Used Python pandas to create correlation charts. As python is one best programming language for data analysts, I have used it to my advantage. Correlation charts in pandas produce heatmap and it is easier to visualize.
3. Used Tableau to create dashboards and stories for presentation. Tableau is a visualization software and its ability to process large data within a few seconds is what made me choose this software.

Findings

Descriptive Statistics:

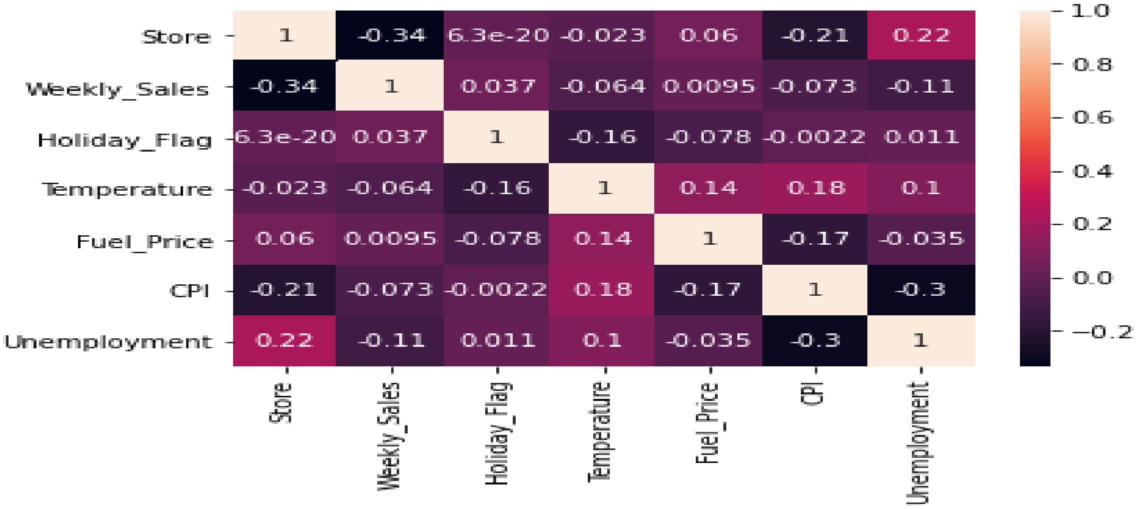
Standard Deviation:

From the table below, we can see that the standard deviation is 564366.62 which implies that there is a lot of variances in the sales among the stores. That means some stores are underperforming.

Statistical Measure	***WeeklySales***	***Temperature***	***Fuel_Price***	***CPI***	***Unemployment***
Mean	1046964.878	60.66378244	3.358607	171.5784	7.999151049
Standard Error	7035.371661	0.229933793	0.005722	0.490619	0.023384704
Median	960746.04	62.67	3.445	182.6165	7.874
Mode	#N/A	50.43	3.638	126.4421	8.099
Standard Deviation	564366.6221	18.44493288	0.45902	39.35671	1.875884782
Sample Variance	3.1851E+11	340.2155488	0.210699	1548.951	3.518943715
Kurtosis	0.053140927	-0.61280096	-1.17738	-1.83981	2.639711784
Skewness	0.668361797	-0.3367676	-0.09616	0.063492	1.188143933
Range	3608700.2	102.2	1.996	101.1688	10.434
Minimum	209986.25	-2.06	2.472	126.064	3.879
Maximum	3818686.45	100.14	4.468	227.2328	14.313
Sum	6737218987	390371.44	21612.64	1104107	51474.537
Count	6435	6435	6435	6435	6435

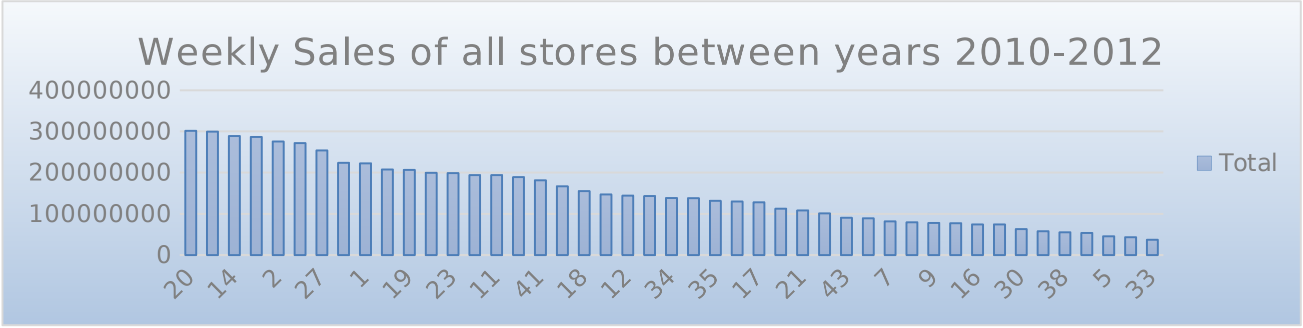
Correlation:

There is no definite relation that weekly sales are impacted by the factors provided in the data. However, there is a negative correlation coefficient for Temperature, CPI and Unemployment which implies that as they are increasing in value sales are decreasing.



Sales Analysis: Impact of holidays on Sales

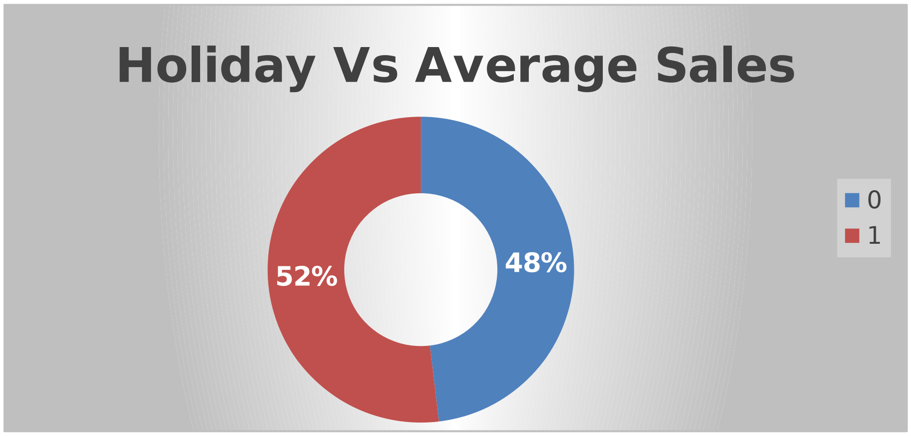
1. From the below bar chart, it is evident that store number 20 has the most sales and store number 33 has the least sales.



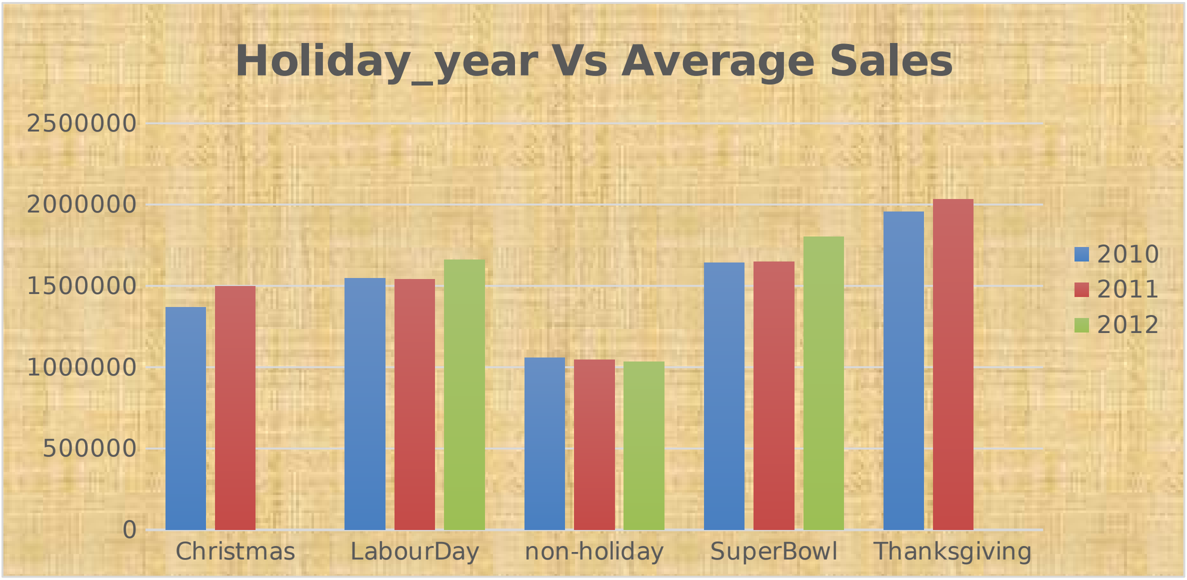
2. In 2011, the sales were highest with $2.54B followed by 2010 with $2.29B all stores combined.

3. The average sales on holidays and events were greater than non-holiday sales.

Row Labels	Average of Weekly_Sales
0	1041256.38
1	1122887.892

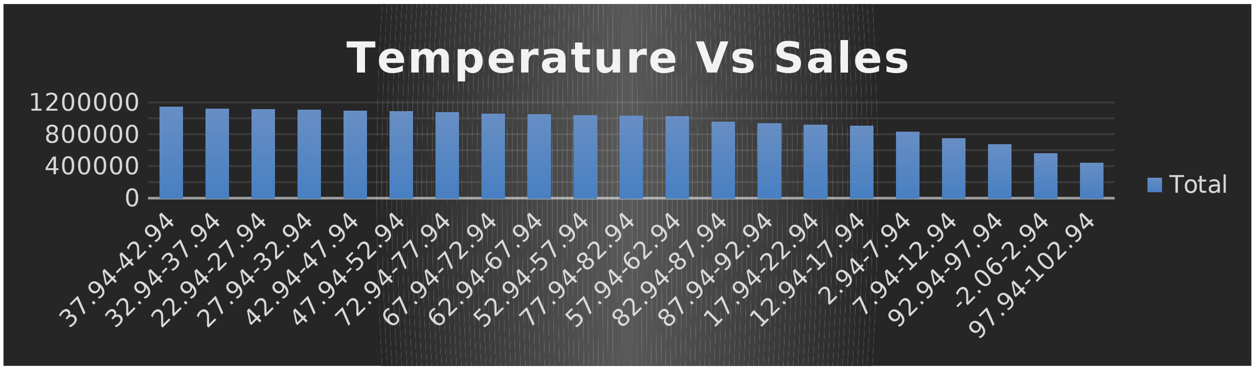


4. Thanksgiving average sales were highest than any other holidays and non-holidays.

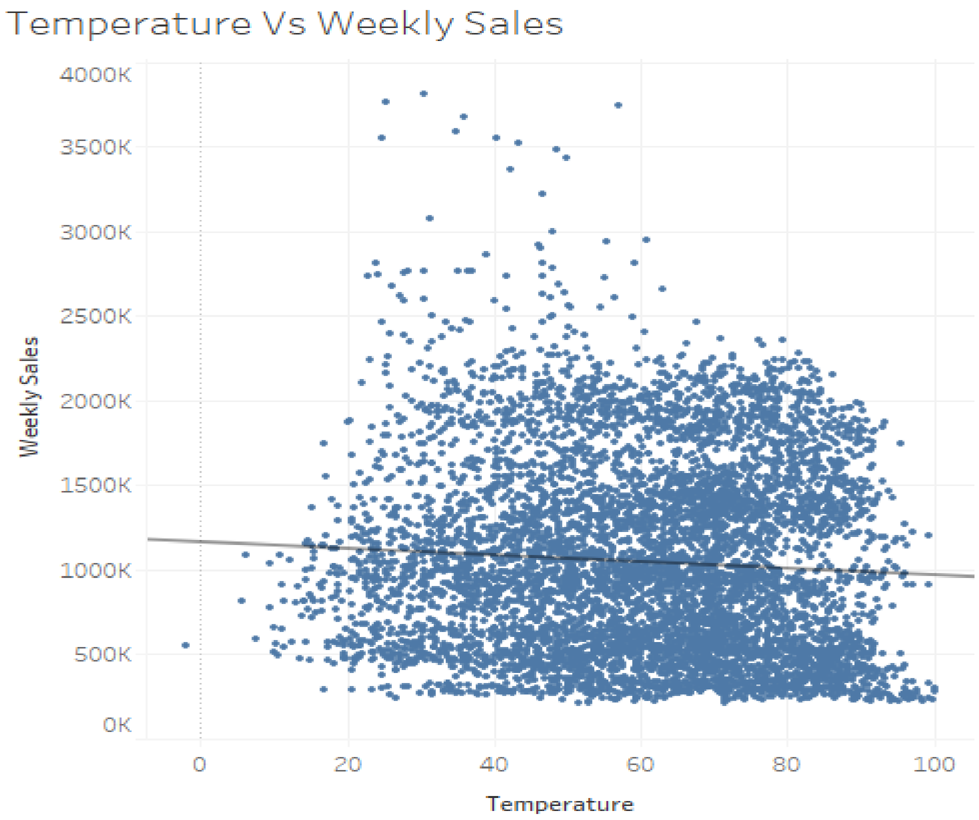


Impact of Temperature

1. The average weekly sales dropped when temperatures are very high and very low. People don’t like to shop when the climates are extreme.

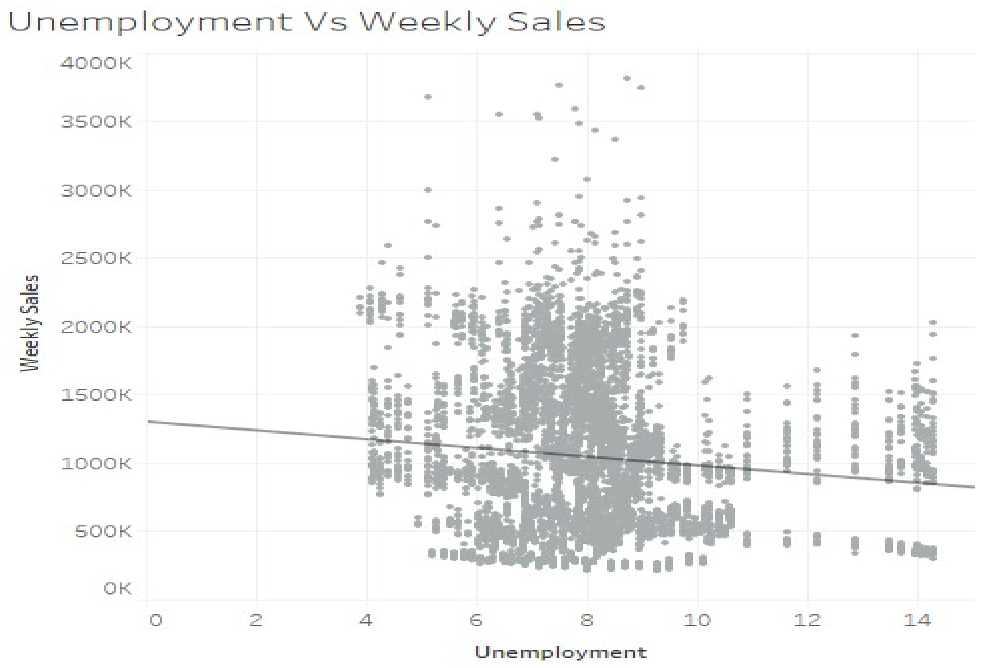


2. From the below scatter plot, we can see that there is a negative trend line which implies that as temperature increases sales are decreasing. Even though there is a slight correlation, it doesn’t imply that it the direct impact on sales.

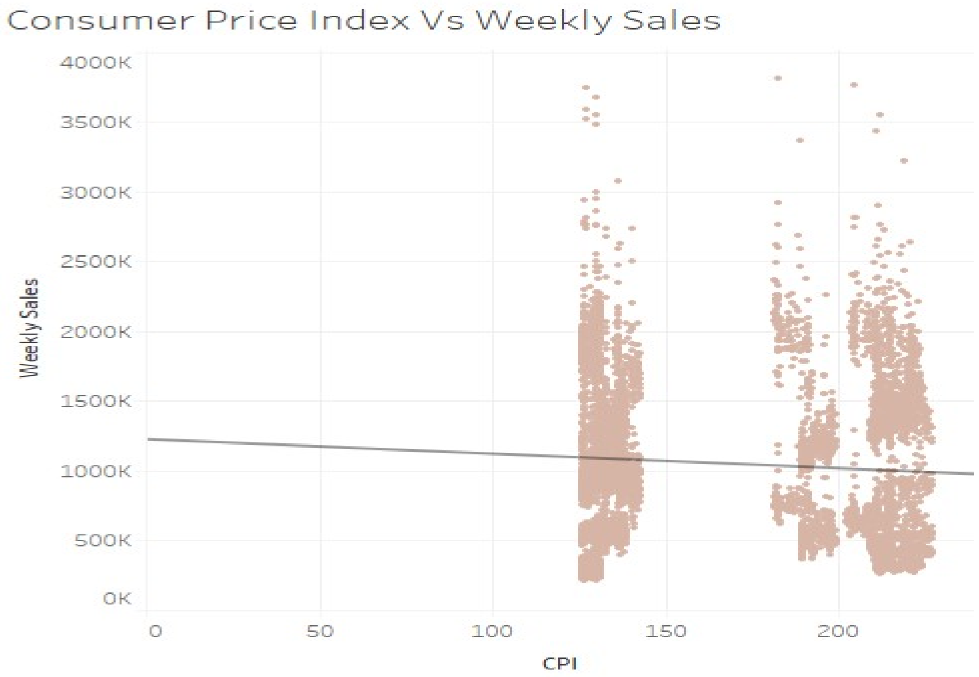


Impact of Unemployment and CPI:

1. As unemployment rate increases sales decreases. The negative trend line in the graph proves the point.

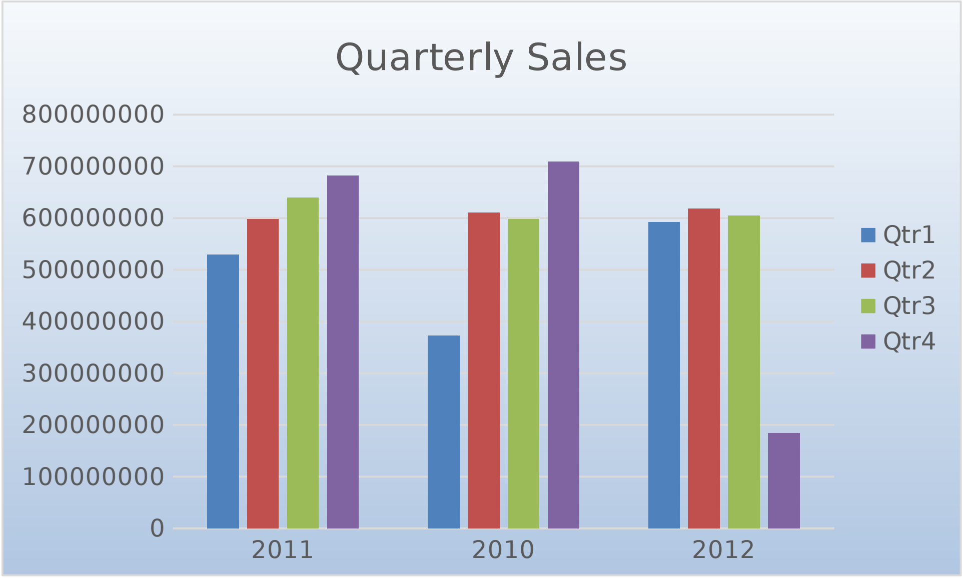


2. Average sales are high when CPI is in medium range. As CPI increases Sales decreases.

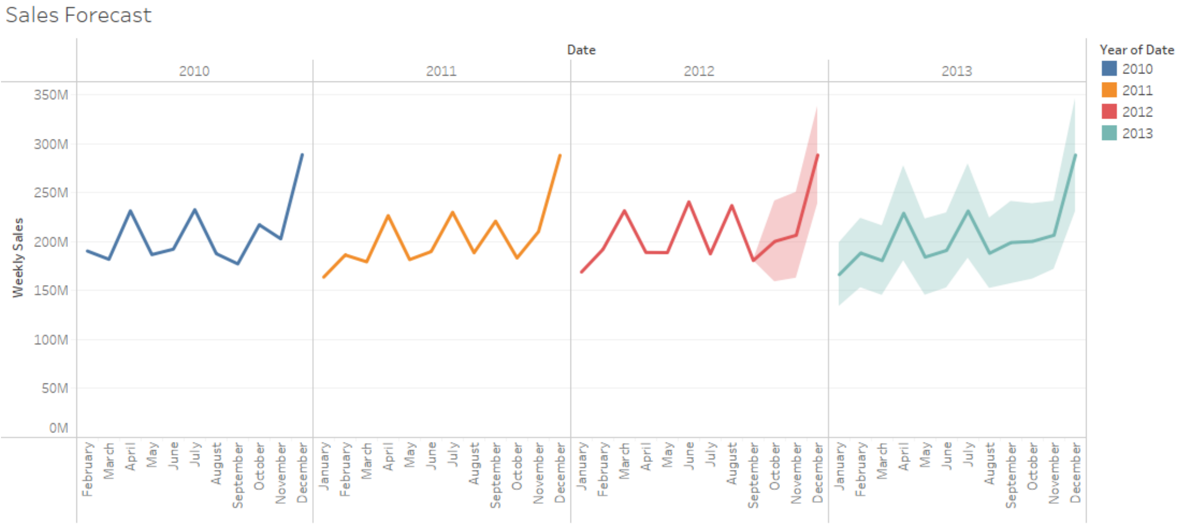


Sales Forecast:

1. From the analysis, we can say that although average sales are high during Thanksgiving, the sales are always high in December. Which says that temperature is not a factor that impacts sales as December has extreme cold weather. Also, quarter 4 have the most recorded sales in every year.



2. From the forecast chart, it is predicted that the sales will be higher in the year 2012.



Data Analysis

Insights:

1. From all the findings, the sales are highest during thanksgiving and in the month of December. So, to fulfil customer needs, stock the stores with enough supply during this period and every holiday event.
2. Some stores are generating less revenue than the average of all stores. Find out what is the problem with those stores like poor customer satisfaction or bad location. If it is a bad location, then the organization needs to relocate the store.
3. As the sales on non-holiday events are low, conduct sudden surprise events in the store and make customers participate in them. This will make the customers happy and excited, and they will be visiting the store more often as there is a surprising element added with the shopping experience.
4. Conduct surveys on what kind of items do customers regularly buy. Also find out what items they want that are not available in the store.
5. To attract more customers during high unemployment rate, slash the prices of necessary items used for the survival of livelihood. This will make the customers think that the organization cares for the needs of the people. It will further enhance the trust in the organization and adds more customers.

Conclusion

The data analysis for the Walmart retail dataset is successfully completed. All the data given is cleaned, sorted, filtered, and organized for analysis. From the analysis, it can be concluded that the store needs to stock more during the holiday events and in the month of December.

Appendix

Excel Sheets Data Set for Udemy Project

https://docs.google.com/spreadsheets/d/1PfECyAWADB3yQ0jBFiNK2g_AqNW4Y71FuQHkqtxLF3Y/edit#gid=1774804965

Tableau visualization for Udemy Project

Udemy_Dashboard | Tableau Public

Excel Sheets Data for Walmart Retail Analysis

https://docs.google.com/spreadsheets/d/1zHq_QOf4zQHfrqo3pxUuwt7eyjJVCPcPbZXFuaCldc/edit#gid=792520748

Tableau visualization for Walmart Retail Analysis

Walmart_retail_analysis | Tableau Public