# Contents

# Introduction

[1]*"Data mining is the process of understanding data through cleaning raw data, finding patterns, creating models, and testing those models."* (www.tableau.com, 2021). The Cross Industry Standard Process for Data Mining phases helps businesses to make accurate predictions, perform forecasts, and recognise outliers.

1. **Business understanding**

Companies offering cloud computing services such as VMware, Microsoft, VMware, Cisco, Red Hat, and Amazon, sell products at various prices mostly through licensing. Those technological pioneers had to face the Brexit impact on their internal operations. Businesses aim to offer consistent Local Currency pricing when possible to provide stability for partners and end customers. (www.vmware.com, 2021). As the United Kingdom decided to leave the European Union in June 2016, British clients should be considered as a sensitive topic as definitely there are some changes to appear over time between the seller and British buyer.

*"Tech companies remain worried about Brexit's impact: Three-quarters of UK tech workers said they believe the business environment may get worse…"* (Rayome, 2016) Moreover, companies such as Apple, Dell, and HP were forced to raise prices for its customers in the United Kingdom as a result of the decrease of the value of GBP. (Rayome, 2016). The goal of cloud computing companies is to have consistent Local Currency pricing when possible, to provide stability for partners and end customers. Issues have arisen when GBP List Prices differ too much from EUR List Prices as the data set indicates.

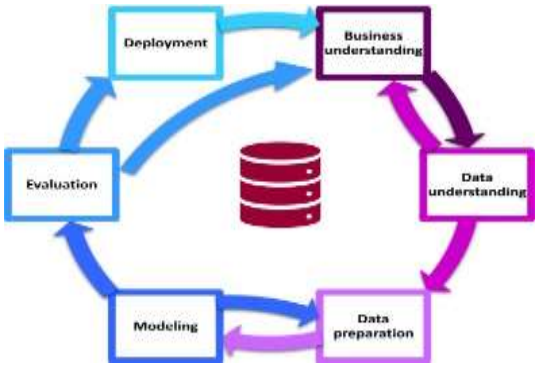


*Figure 1. 1Cross Industry Standard Process for Data Mining*

**Data understanding**

Considering Brexit and the company's pricing strategies, the data about product pricing and monetization should be used for data analysis.

[1] https://ars.els-cdn.com/content/image/3-s2.0-B9780128191545000102-gr003.jpg

# Body

## Objectives

1. What is the average price discount to the clients? How many clients receive more or less than the average discount rate?
2. How Brexit has impacted the cloud computing providers 'product prices for British customers? Have these List Prices decreased or increased? Is there a difference between List Prices for the United Kingdom, France, and Germany?
3. Is there any correlation between the discount rates for customers in the United Kingdom, France, or Germany and the prices of products?
4. What market is assigned with the lowest discounts?

## Dataset description – summary of the data

I have been using the dataset from my workplace with full permission to use this data not mentioning the name of the company would secure the data itself. The dataset represents the list prices for 16 products, which are owned by the cloud computing company. The table represents the discounting strategy as one of the strongest pricing operations in the B2B environment, which have been used in 2016 with customers from the United Kingdom, Germany, and France. The table consists of 11 columns including Average Selling Price, quantities, products, discounts and the targeted country. However, our interest lies in those labelled Country, Product, Currency, List Price USD, and Discount %. Furthermore, the table was subtracted into 3 columns only for the analysis purposes.

**Number of columns and rows**

The data set was aligned to our requirements and the number of columns decreased by creating a new variable called new_prices_UK. This dataset represents 3 columns and 18780 rows, which were used for data analysis and findings.

The columns in the data set are labelled accordingly:

Column 1: Product

Column 2: LP USD

Column 3: Discount %

```
#Filtering columns + removing missing values from it
new_prices_uk = prices_uk[['Product','LP USD', 'Discount %']].dropna()
print(type(new_prices_uk))
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
#Overview of a substracted dataset

print(new_prices_uk.info())
new_prices_uk.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 18780 entries, 0 to 18779
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Product     18780 non-null  object
 1   LP USD      18780 non-null  float64
 2   Discount %  18780 non-null  float64
dtypes: float64(2), object(1)
memory usage: 586.9+ KB
None
```

| | Product | LP USD | Discount % |
|---|---|---|---|
| 0 | Product A | 272.151899 | 0.529231 |
| 1 | Product A | 272.151899 | 0.529231 |

**Detail data types**

Data analysis has used different Data types including arrays, DataFrames, a dictionary, 4 lists, Series, floats, strings, and tuples.

## Data pre-processing

```
# In[207]:
#Step 1. Data preprocessing

#DataFrame subsetting to see prices in GBP only
prices_uk = prices[prices['Country'] == 'UK']
```

```
# In[208]:
#Printing out the first 5 rows of data in my dataset
prices_uk.head()
#Checking number of rows and columns of the new dataset
prices_uk.shape
```

```
# In[209]:
#Filtering columns + removing missing values from it
new_prices_uk = prices_uk[['Product','LP USD', 'Discount %']].dropna()
print(type(new_prices_uk))
```

```
# In[210]:
#Overview of a substracted dataset
print(new_prices_uk.info())
new_prices_uk.head()
```

Data analysis began with calculating weighted average ratings to aggregate the descriptive statistics including mean, standard deviation, minimum value, and maximum value. The output shows the statistics for two columns of our interest.

**Descriptive statistics**

| | List Price $ | Discount % |
|---|---|---|
| count | 18780.000000 | 18780.000000 |
| mean | 642.924738 | 0.512992 |
| std | 444.307248 | 0.110190 |
| min | 272.151899 | 0.150000 |
| 25% | 322.784810 | 0.473565 |
| 50% | 322.784810 | 0.524114 |
| 75% | 1132.911392 | 0.576555 |
| max | 2094.936709 | 0.800000 |

Besides values received after using describe the function, additional functions were used to establish the median, mode, and skewness of the List Prices. If the mean is lower than the median and mode it was established that the skewness is negative.

Calculating skewness allowed us to understand the direction of outliers and allow us to create a linear model.

```
Name: List Price $, Length: 18780, dtype: float64
Mean of the List prices is set to: 642.9247381404941
Mode of the List prices is set to: 0    322.78481
dtype: float64
Median of the List prices is set to: 322.7848101265823
The skewness of the list prices is set to: 0.8536965121956172
```

## Data analysis

### Insight 1

**What is the average price discount to the clients? How many clients receive more or less than the average discount rate?**

```
In [83]: #Converting a float to percentage value
         discount_average = "{:.0%}".format(discount_rate.mean())
         print("Average price discount is estimated to be " + str(discount_average) + " .")

         Average price discount is estimated to be 51% .
```

The average price discount assigned to the customers is equal to 51%, and it was established that the range excluding outliers lies between 32% and 73%, meaning that the majority of clients receive a discount lying in this range.

Assigning number of outliers allow to establish the data points which are distant from all other values. *"Data point that falls outside of 1.5 times of an interquartile range above the 3rd quartile and below the 1st quartile. Data point that falls outside of 3 standard deviations. we can use a z score and if the z score falls outside of 2 standard deviation" (Khandelwal, 2018).*

If the outlier becomes identified, it can prevent incorrect calculation of skewness, mean and standard deviation. It is safer to check the outliers before arranging the descriptive statistics. After using IQR to see how spread the middle values are, it was established that the outliers are lower than 32% and higher than 73% and the total number of outliers is equal to 1159, meaning that 1159 customers receive an unusual discount rate, which is assigned to them under some extraordinary conditions.

**5 Steps to calculate the number of outliers for the Discount % column**

In [22]:
```
#Calculating outliers for a Discount % column

#Step 1: Sorting the 'Discount %' column values in ascending order

discount_sorted = new_prices_uk.sort_values(["Discount_%_UK"], ascending = True)
print(discount_sorted)
```

In [23]:
```
#Step 2: Calculating IQR value
discount = discount_sorted['Discount_%_UK']
q1 = discount.quantile(0.25)
q3 = discount.quantile(0.75)
iqr = q3 - q1
print(iqr)


#https://www.kite.com/python/answers/how-to-format-a-number-as-a-percentage-in-python
#Converting a float to percentage value
iqr_percentage = "{:.0%}".format(iqr)
print("The IQR value is equal to " + str(iqr_percentage) + " .")
```

In [106]:
```
#https://medium.datadriveninvestor.com/finding-outliers-in-dataset-using-python-efc3fce6ce32 Renu Khandelwal
#Step 3: Calculating lower and upper bounds
lower_bound = q1 - (1.5 * iqr)
upper_bound = q3 +(1.5* iqr)

#Converting a float to percentage value
lower_bound_percentage = "{:.0%}".format(lower_bound)

upper_bound_percentage = "{:.0%}".format(upper_bound)
print("The range excluding outliers lies between " + str(lower_bound_percentage) +
      " and " + str(upper_bound_percentage) + " .")

#Step 4: Calculating discount outliers
discount_outliers = discount_sorted[(discount < q1 - 1.5 * iqr) | (discount > q3 + 1.5 * iqr)]
print(discount_outliers)
```

In [26]:
```
#Step 5: Calculating number of outliers in the Discount % column
print("Outliers include all the values lower than " + str(lower_bound_percentage) +
      " and higher than  " + str(upper_bound_percentage) +" and the total number of outliers is equal to : "
     + str(discount_outliers.shape[0] + int(1)) + " .")
```

```
Outliers include all the values lower than 32% and higher than  73% and the total number of outliers is equal to : 1159 .
```

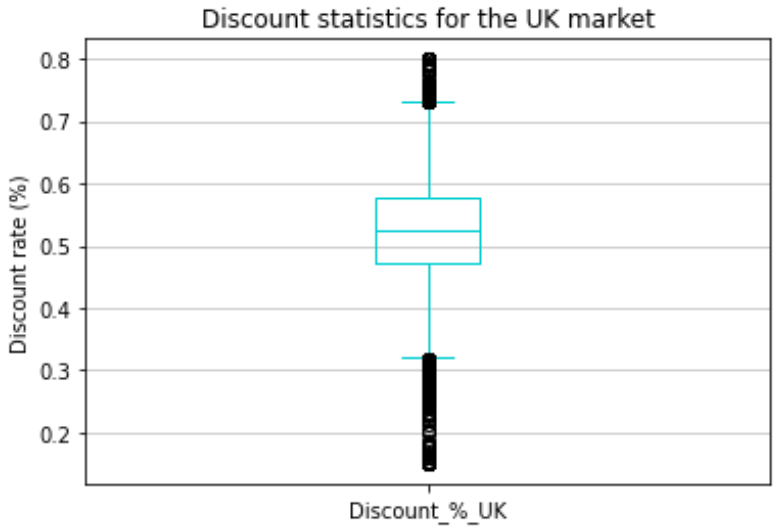



*Figure 1. Box plot displaying outliers for the Discount rates for the UK market.*

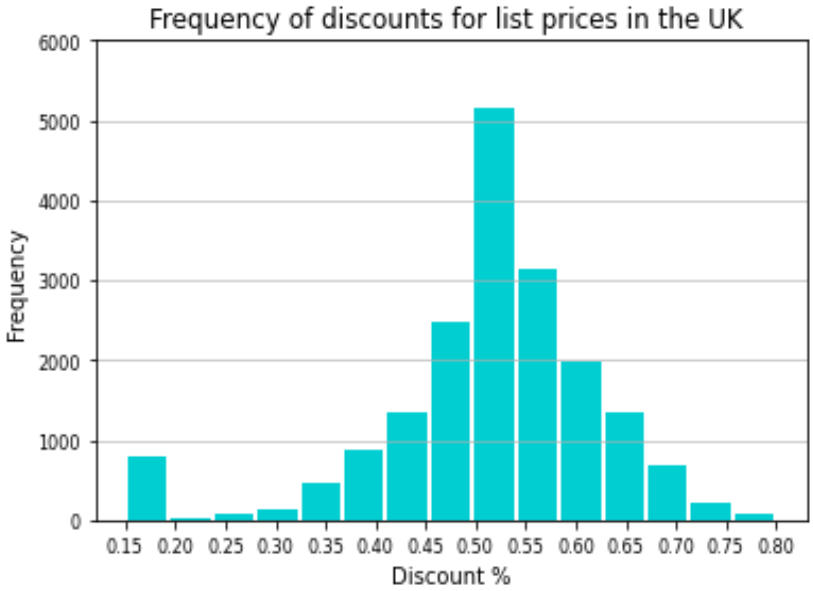



*Figure 2. Histogram representing the frequency of discounts for list prices in the UK*

The histogram was used to represent a continuous data set, simultaneously it displays the outliers and skewness and the number of occurrences of scores within each individual bin. The bin width was chosen to have 5% difference between one interval and another, starting from 15% and ending at 80%. A frequent discount of 51% is assigned for the majority of products' prices in the United Kingdom, it can be estimated that over 5100 customers receive the highest discount rates. Also, almost 900 customers receive only a 15% discount, which is a quite high number of contracts signed compared to the number of customers receiving 20%, 25%, 30%, 35%, 70%, 75%, or 80% discounts. Furthermore, the bars represent the range of the highest discount rates from 32% to approximately 70%.

Insight 2

**How Brexit has impacted the cloud computing providers' product prices for British customers? Have these List Prices decreased or increased? Is there a difference between List Prices for the United Kingdom, France, and Germany?**

Establishing average list prices for all 16 products in EUR and GBP helped to establish the difference in prices accordingly to the change of conversion rate for GBP currency from USD. (decrease from 0.79 to 0.75).

| Oct 1 2016 | EUR | GBP |
|---|---|---|
| USD | 0.88 | 0.79 |
| **Dec 7 2021** | **EUR** | **GBP** |
| USD | 0.89 | 0.75 |

Table 1. Conversion rates for EUR and GBP from USD

After conducting data analysis, after calculating the list prices for all the 16 products converting euro to dollars and pounds to dollars it has been discovered that the list prices of all the products converted from GBP to USD are lower than the prices converted from EUR. It was caused by the movement in Local Currency prices as Brexit caused the pound value to decrease in the advantage of the USD by 20%.

It is recommended to increase the GBP List prices for all the products according to the difference in the list prices between the List Price in $ from EUR to the List Price in $ from GBP. Prices should be equal for all British, French, and German customers. It has been calculated to increase the prices for cloud computing services adjust the prices and be more currency movement aware. It can be seen from Table 1 how the GBP value has dropped from 2016 to December 2021 observing the conversion rate decrease for GBP, from 0.79 to 0.75.

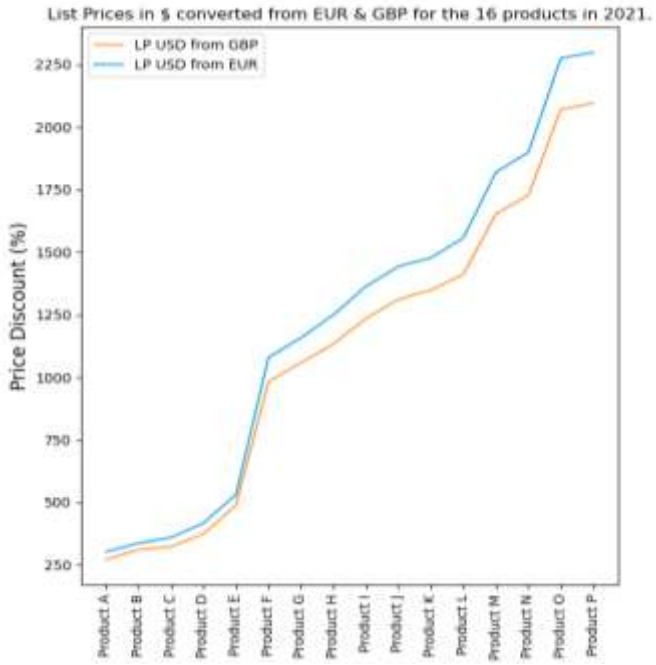



*Figure 2. The line chart displaying the difference between the List Prices in 2021 for 16 products.*

Insight 3

**Is there any correlation between the discount rates for customers in the United Kingdom, France or Germany and the prices of products?**

To establish changes in the List Prices converted from EUR and from GBP, there were used correlation and linear regression methods.

First of all, the scatterplots were used to display the correlation between the List Prices from GBP vs Price Discount rates for the United Kingdom.

```
scatterplot.plot(kind = 'scatter',
                 x = 'List Price(UK)',
                 y = 'Discount %(UK)',
                 title = "List prices vs Discount rates for the UK customers",
                 color = '#14AAF5',
                 yticks = [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0],
                 xticks = [0,200,400,600,800,1000,1200,1400,1600,1800,2000,2200,2400],
                 fontsize = 8,
                 edgecolor = 'black',
                 linewidth = 1,
                 alpha = 0.75
                )

plt.grid(axis = 'y', alpha = 0.75)
plt.show()
```

The code above was used to draw a scatterplot, which illustrates the relationship between 2 continuous numerical fields.

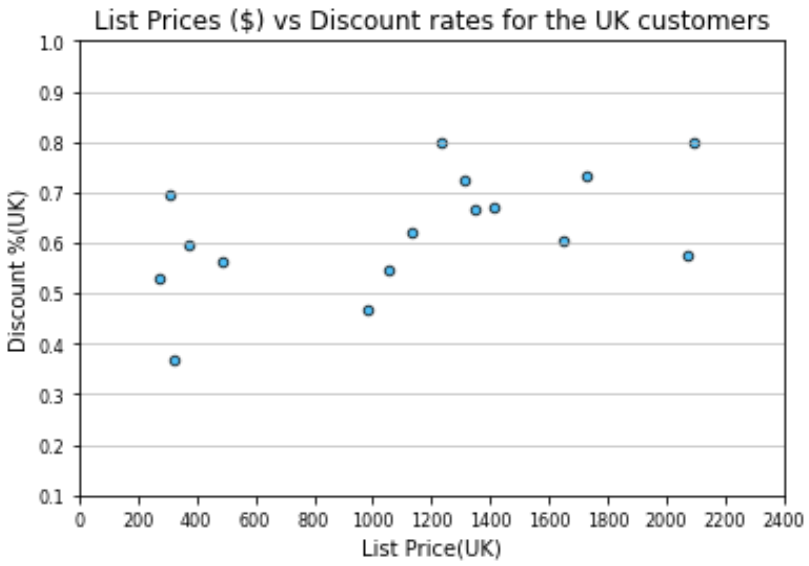



*Figure 4. The scatterplot displays the relationship between the List Prices in $ converted from GBP prices and Discount rates (%) for the UK customers.*

The second scatterplot represents the relationship between the List Prices ($) converted from EUR vs Discount rates for France and Germany. It required dividing one dataset into two separate data frames and then dividing the second one into two separate to gain only values for France and separate values for Germany.

```
#DataFrames preprocessing
#DataFrame subsetting to see prices for French customers
prices_france = prices[prices['Country'] == 'France']
print(prices_france)
```

```
#Filtering columns
new_prices_france = prices_france[['Product','LP USD', 'Discount %']]
print(type(new_prices_france))
print(new_prices_france)
```

```
#DataFrame subsetting to see prices for German customers
prices_germany = prices[prices['Country'] == 'Germany']
print(prices_germany)
```

```
#Filtering columns
new_prices_germany = prices_germany[['Product','LP USD', 'Discount %']]
print(type(new_prices_germany))
print(new_prices_germany)
```

Furthermore, two separate DataFrames were merged by rows into one indicating customers in both Germany and France who paid for the products in EUR.

```
#Merging two dataframes by rows into one indicating customers in Germany and France who pay in EUR currency
prices_eur = pd.concat([new_prices_france, new_prices_germany], axis = 0)
print(prices_eur)
```

```
         Product  List Price $  Discount %
35633  Product A        302.20        0.35
35634  Product A        302.20        0.18
35635  Product A        302.20        0.18
35636  Product A        302.20        0.15
35637  Product A        302.20        0.15
...          ...           ...         ...
35628  Product P       2297.87        0.44
35629  Product P       2297.87        0.43
35630  Product P       2297.87        0.46
35631  Product P       2297.87        0.43
35632  Product P       2297.87        0.34

[32754 rows x 3 columns]
```

As a result, the new data frame consists of 32754 rows as many of these are duplicated by product name. I was able to subset the Data Frame to see only the prices and discount rates for each of 16 products for Germany and France. The final data set can be used for the data analysis for this third insight.

First of all, the second scatterplot was created to visualise the relationship between the List Prices in $ vs Discount rated for the French and German customers.

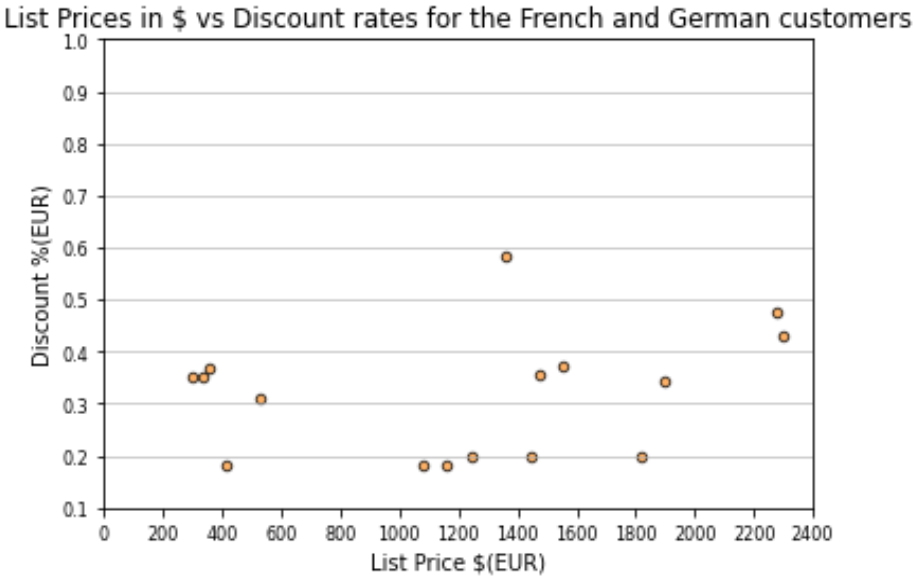



*Figure 5. Scatterplot displaying the relationship between List Prices in USD vs Discount rates for the French and German customers.*

### *Regression test*

Furthermore, comparing two scatterplots would present the difference between both relationships, as a result, the multiple scatterplot was created, which is always useful for interpreting trends in the statistical data. The regression test indicates the cause-and-effect relationships between variables, and it was used to estimate the effect of one variable on another comparing two pairs of variables.

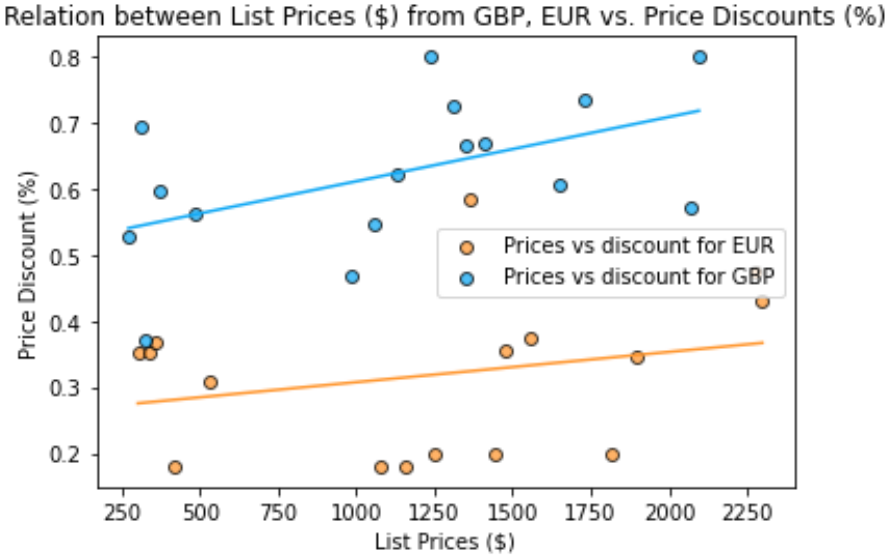



*Figure 3. Scatterplot displaying the regression between the List Prices in USD converted from GBP and EUR vs Discount rates including linear regression.*

The scatterplot indicates the finding, that there is a positive relationship between X and Y and each pair of variables forms a linear relationship. I was able to visually inspect the graph's scape and it can be established that the diagram above displays two linear relationships.

**Findings include:**

- The discounts are lower for local customers in France and Germany compared to the United Kingdom.
- Lower profits for the company from the British customers, who pay in GBP.
- Positive relationship between X and Y.
- The higher the List Prices set for the products; the higher the Price Discounts assigned to the customers for all the countries paying in GBP and EUR.

### *The Pearson Correlation test*

This type of test was used to check whether List Prices converted from two different currencies into USD have relationships with Discount rates set for 16 different products in countries such as Germany, France, and the United Kingdom. The Pearson correlation coefficient measures the linear relationships between two datasets.

### *Correlation test 1*

For calculating correlation can be used two libraries, **Pandas** or **NumPy**.

```
In [156]: #Pearson correlation test
          #https://www.youtube.com/watch?v=TRNaMGkdn-A  author: stikpet


          #Correlation for list prices vs discounts for the UK
          scatterplot_dataset_1.corr()
```

Out[156]:

| | List_Price_UK | Discount_%_UK |
|---|---|---|
| List_Price_UK | 1.00 | 0.51 |
| Discount_%_UK | 0.51 | 1.00 |

A correlation coefficient between the UK List Prices and Discounts is equal to 0.51, which is **positive**, meaning that if the List Price is then the customer usually receives the higher discount. The strengths of the association are large, as it is greater than 0.5.

**r = 0.51**

0.51 > 0 ---- > Positive Correlation

0.5 to 1.0 --- > Large strength

*p-value test 1*

```
In [166]: #Calculating the p-value (significance) for the UK
          #Importing pearsonr function from SciPy
          from scipy.stats import pearsonr

          pearson = pearsonr(scatterplot_dataset_1['List_Price_UK'], scatterplot_dataset_1['Discount_%_UK'])

In [165]: print('The p-value is equal to ' + str(pearson[1]) + '.')

          The p-value is equal to 0.041986920505568785.
```

**p-value = 0.04**

There is a 4% risk of concluding that there is a difference between the variables.

**Null hypothesis:** No correlation between the List Price for the UK and the Discount rate

**Alternative hypothesis:** Correlation between the List Price for the UK and discount rate

**p-value < 0.05**

The p-value is less than the significance level so we can reject the Null hypothesis, however, there is only moderate evidence against the null hypothesis in favour of the alternative, meaning that there is a correlation between the List Price for the UK and Discount rate.

In theory, p-value determines the probability of extreme results on the statistical hypothesis test, taking the Null Hypothesis to be correct.

*If the p-value is below 0.5, we would consider it significant.

*Correlation test 2*

```
In [157]: #Correlation for list prices vs discounts for the France and Germany
          scatterplot_dataset_2.corr()
```

Out[157]:

| | List_Price_EUR | Discount_%_EUR |
|---|---|---|
| List_Price_EUR | 1.00 | 0.26 |
| Discount_%_EUR | 0.26 | 1.00 |

A correlation coefficient between the List Prices and Discounts for France and Germany is equal to 0.26, which is **positive**. The higher the List Price is, the customer usually receives the higher discount, however in this case the strengths of the association are small, as it lies between the range of 0.1 and 0.3.

*p-value test 2*

```
In [171]: #Calculating the p-value (significance) for the France and Germany
          pearson2 = pearsonr(scatterplot_dataset_2['List_Price_EUR'], scatterplot_dataset_2['Discount_%_EUR'])
          print('The p-value is equal to ' + str(pearson2[1]) + '.')

          The p-value is equal to 0.3360777382541876.
```

**p-value = 0.33**

There is a 33% risk of concluding that there is a difference between the variables.

**Null hypothesis:** No correlation between the List Price for France and Germany and the Discount rate.

**Alternative hypothesis:** Correlation between the List Price for France and Germany and Discount Rate.

**p-value >0.05**

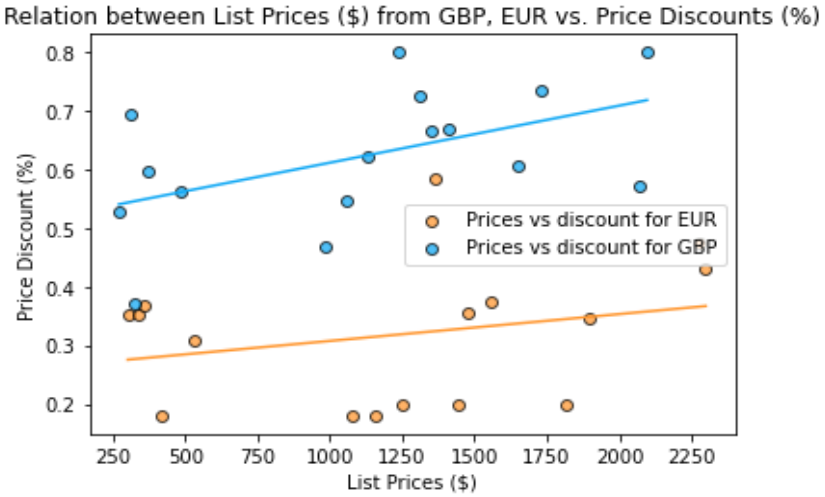
The p-value is greater than the significance level so we can accept the Null Hypothesis and reject the Alternative Hypothesis, meaning that **there is no correlation** between the List Price for France and Germany and the Discount rate.

In theory, the p-value determines the probability of extreme results to the statistical hypothesis test, assuming the Null Hypothesis to be correct.

## Insight 4

**What market is assigned with the lowest discounts?**

It can be assumed that the blue trendline for the independent variable of the List Price from the customers paying locally in GBP has been higher in the diagram than the orange trendline which indicates the higher price discounts. Comparing the Local List prices in this case, converted to USD, the discounts for British customers are higher than those for French and Germans who pay in EUR currency. However, I was able to establish what country is assigned to the lowest discounts.




*ANOVA test*

H1: μ1=μ2=μ3 ----- > There is no difference in the mean of discount rates for all the countries

Ha: μ1≠μ2≠μ3 ----- > There is a difference in the mean discount rates between Germany, France, and the United Kingdom

The analysis of variance as an ANOVA test indicates the variance between more than two groups. First, the means were calculated for the three groups.

1. Discount rates for Germany – 29% - the lowest
2. Discount rates for France – 38% - the second place
3. Discount rates for the UK – 51% - the third place

```
# In[280]:

#Calculating average price discount rate for the UK
print(discount_rate.mean())
#Calculating average price discount rate for Germany
new_prices_germany_disc = prices_germany['Discount %']
print(new_prices_germany_disc.mean())
#Calculating average price discount rate for France
new_prices_france_disc = prices_france['Discount %']
print(new_prices_france_disc.mean())
```

*Figure 4The code was used to calculate the means of discounts for all three countries.*

The F value is a ratio of the column variance to the error variance.

**F(2) = 19553.3 , p <0.05**

There is a less than 5% chance that there is no difference between the averages of discounts. As a result, we can reject the Null hypothesis, the Alternative hypothesis is true. In conclusion, **there is a significant difference in the discount rates for the three different countries.**

Answering question 4 it can be definitely seen that Germany receives the lowest discounts, whereas the British customers receive the highest discount rates when purchasing products from this cloud computing company.

## Conclusion

The results of the data analysis allowed us to achieve all the objectives and to answer all the questions asked. Although, the data set included over 50 thousand rows, the data itself wasn't complex. Sometimes product prices increase because of some unprecedented circumstances, for instance, Brexit has an impact on the global economy. All the tests of ANOVA, Pearson Correlation, and Regression including the descriptive statistics allowed us to see the differences in discount pricing strategy for three different countries. It is recommended to implement some changes to the List Prices and Discount rates of the cloud computing company after reviewing all the findings. Furthermore, an interesting finding is that **the higher the List Prices set for the products; the higher Price Discounts assigned to the customers in Germany, France, and the United Kingdom** indicates the pattern which could be still kept by the company and could be changed. In my opinion, if the new product will be offered at a higher price, there is no requirement to offer a high discount straight away for the customers, who for instance have a high willingness to pay anyway.

# Bibliography


- Abhigyan (2020). Understanding the P-Value in Regression. Available at: https://medium.com/analytics-vidhya/understanding-the-p-value-in-regression-1fc2cd2568af [Accessed on: 14.12.2021].
- Docs. Scipy.org (2021). Scipy.stats.pearsonr . Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html [Accessed on: 14.12.2021].
- Hoare, T (2016). Business and Data Undestanding. Available at: https://mymoodle.ncirl.ie/pluginfile.php/86645/mod_resource/content/0/Business%20and%20Data%20Understanding.pdf [Accessed on: 14.12.2021].
- Judge, P. (2016). Microsoft hikes UK cloud price after Brexit causes pound to fall. Available at: https://www.datacenterdynamics.com/en/news/microsoft-hikes-uk-cloud-price-after-brexit-causes-pound-to-fall/ [Accessed on: 14.12.2021].
- Khandelwal.R. (2018). Finding outliers in dataset using python. Available at: https://medium.datadriveninvestor.com/finding-outliers-in-dataset-using-python-efc3fce6ce32 [Accessed on: 14.12.2021].
- Laerd statistics. (2021). Pearson Product- Moment Correlation. Available at: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php [Accessed on: 14.12.2021].
- Rayome, A. (2016). Brexit bombshell: Microsoft raises cloud pricing 22% for UK businesses. Available at: https://www.techrepublic.com/article/brexit-bombshell-microsoft-raises-cloud-pricing-22-for-uk-businesses/
- Sharma. A. (2020). Statistics for Data Science: What is Skewness and Why is it Important? Available at: https://www.analyticsvidhya.com/blog/2020/07/what-is-skewness-statistics/ [Accessed on: 14.12.2021].
- Tableau. (2021). How Data Mining Works: A Guide. Available at: https://www.tableau.com/learn/articles/what-is-data-mining [Accessed on: 14.12.2021].
- YouTube, (2020). Python – Pearson Correlation (coefficient and test) Available at: https://www.youtube.com/watch?v=TRNaMGkdn-A [Accessed on: 14.12.2021].