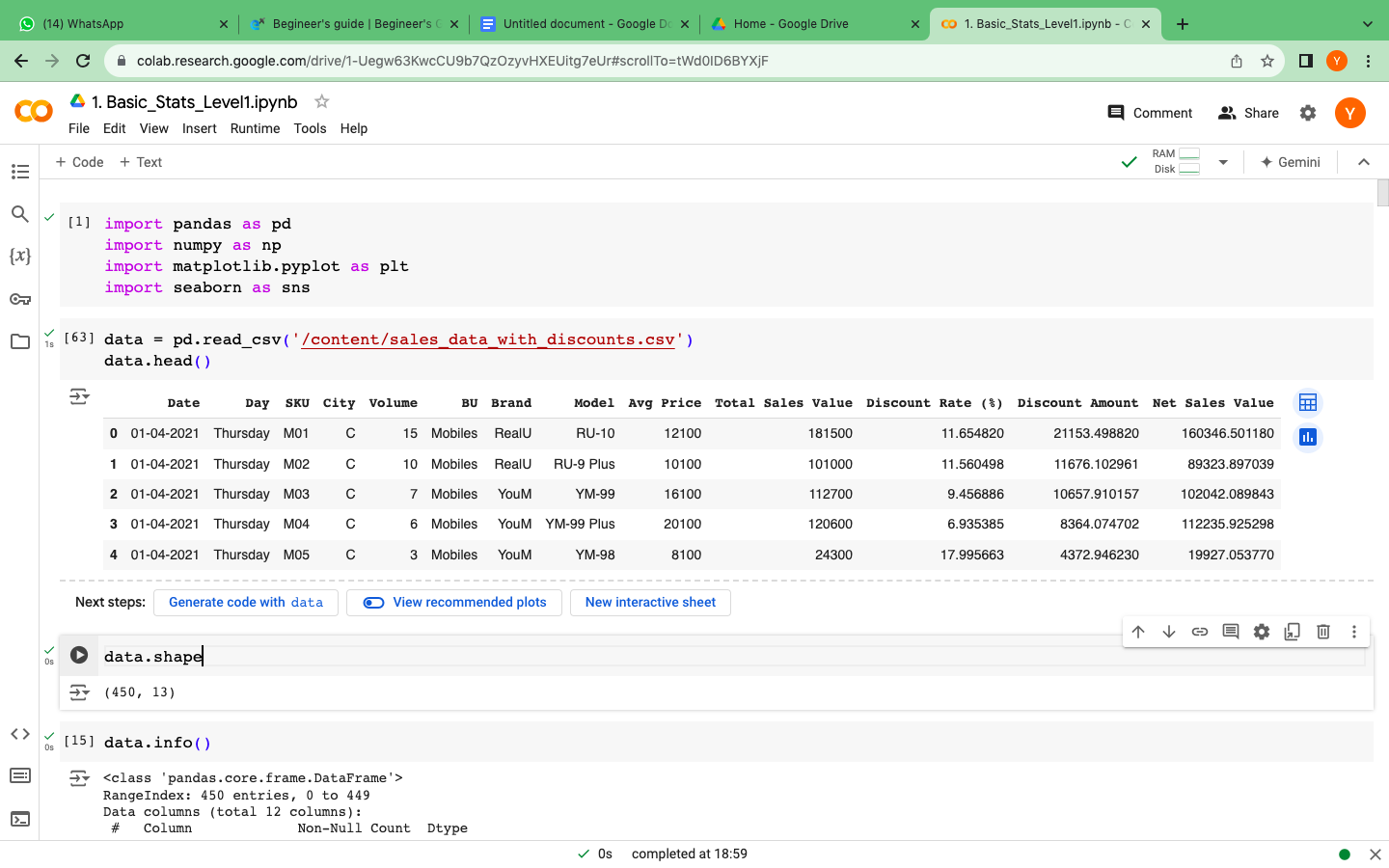
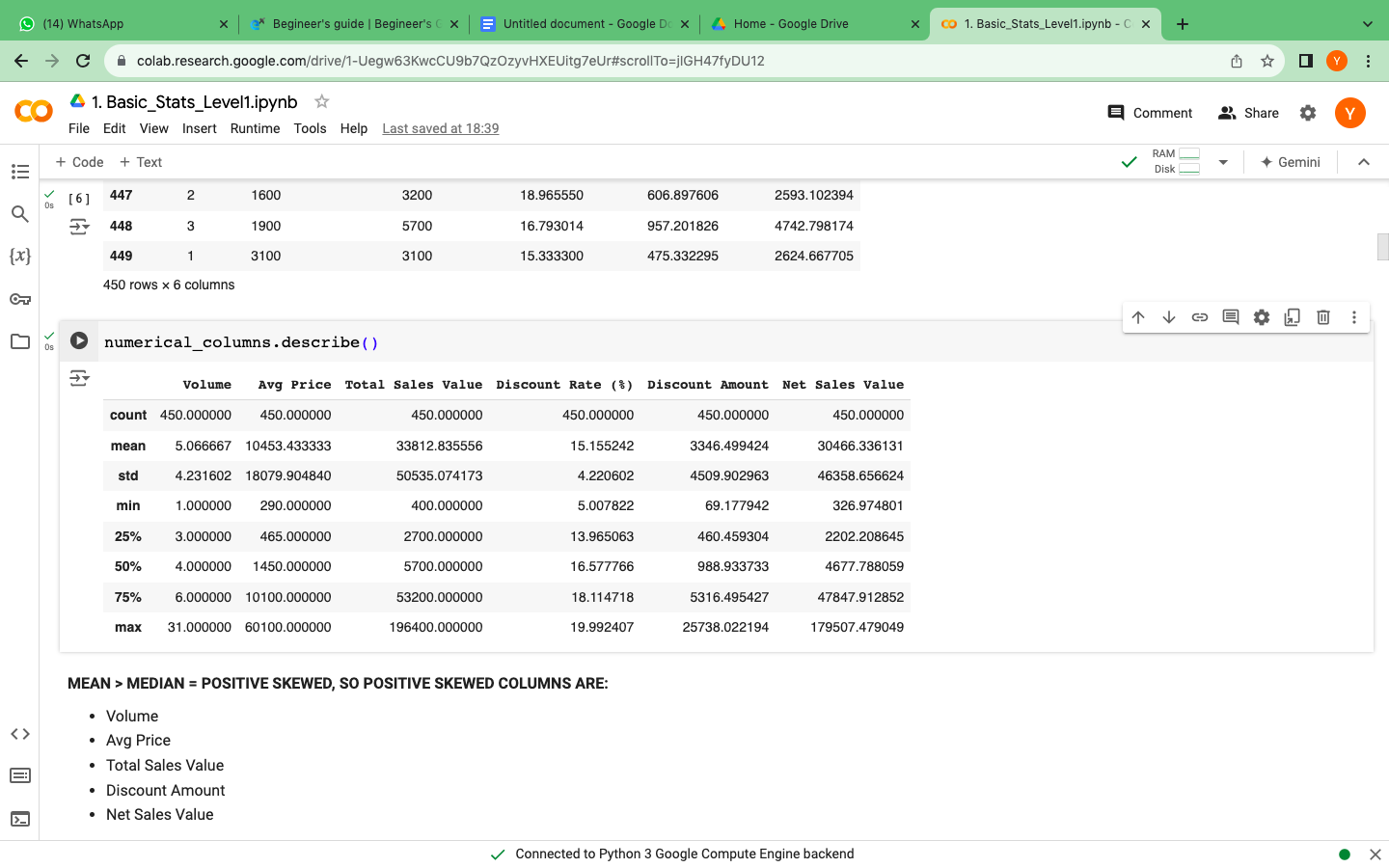
ASSIGNMENT 1: BASIC STATS LEVEL 1

* Descriptive Analytics for Numerical Columns

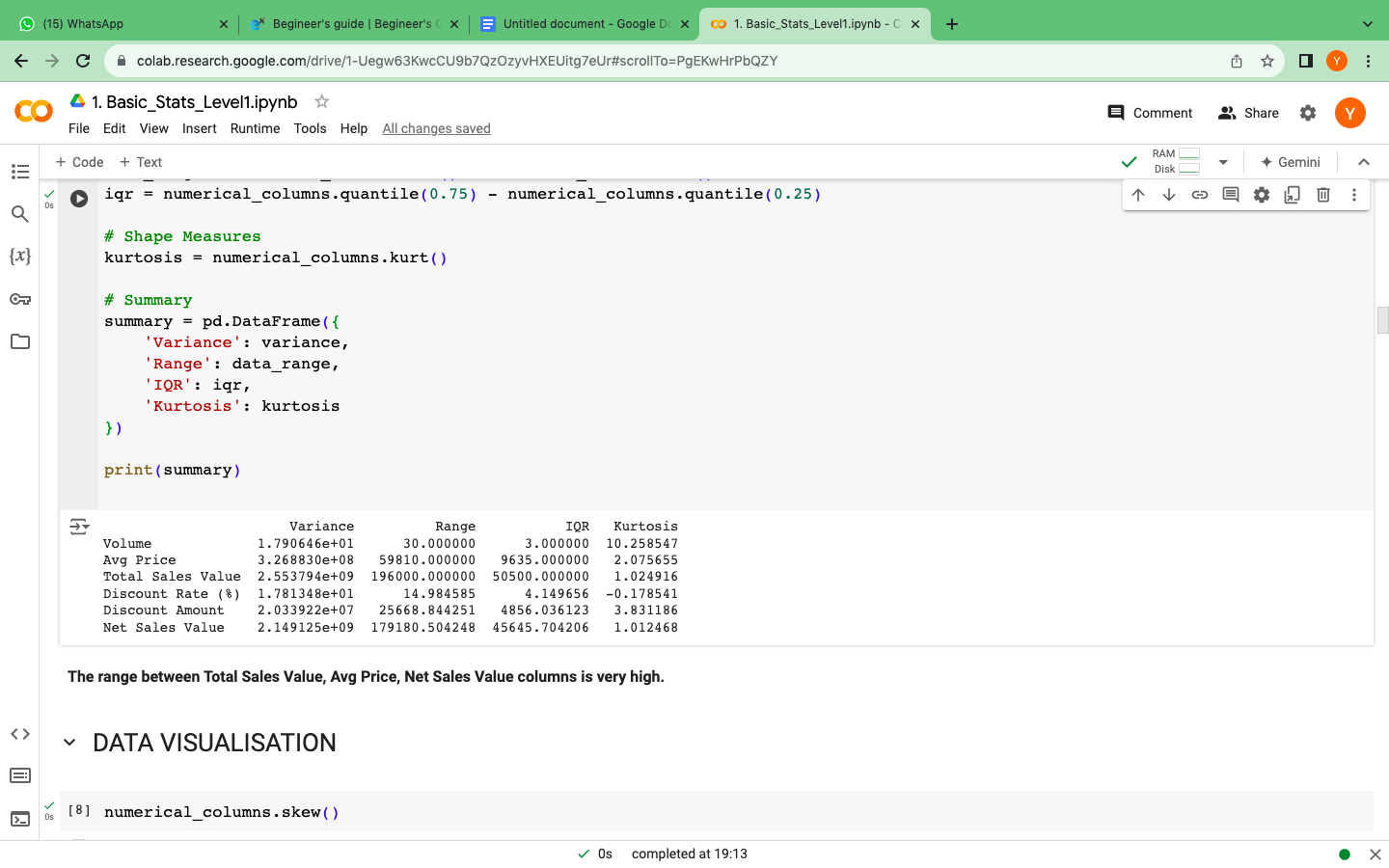


* Our dataset has 450 rows and 13 columns.
* ‘Date’ column was of no use, so we dropped it.
* Has no missing values.
* Out of 12 columns, 6 are categorical and 6 are numerical.

***STATISTICAL MEASURE***



Let's check variance, range and kurtosis



**Range** : The difference between the maximum and minimum values, so columns with high range are:

* Avg price
* Total sales value
* Net sales value
* Discount amount

**Kurtosis** : High kurtosis indicates more outliers and extreme values than a normal distribution, so columns with high kurtosis are:

* Volume
* Discount amount
* Data Visualisation

*Mean > median = positive skewness*

* The columns with positive skewness are :

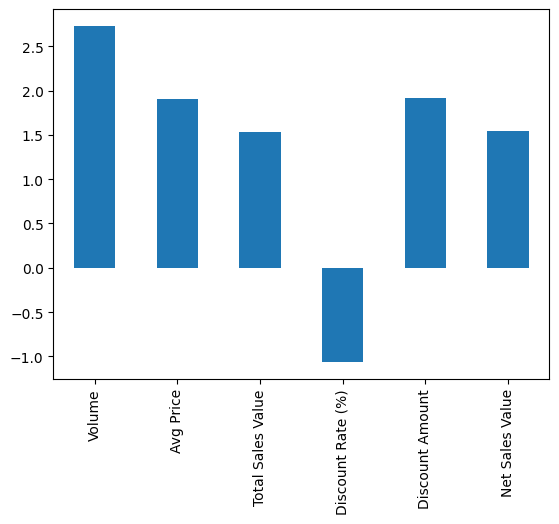
\* Volume

\* Avg Price

\* Total Sales Value

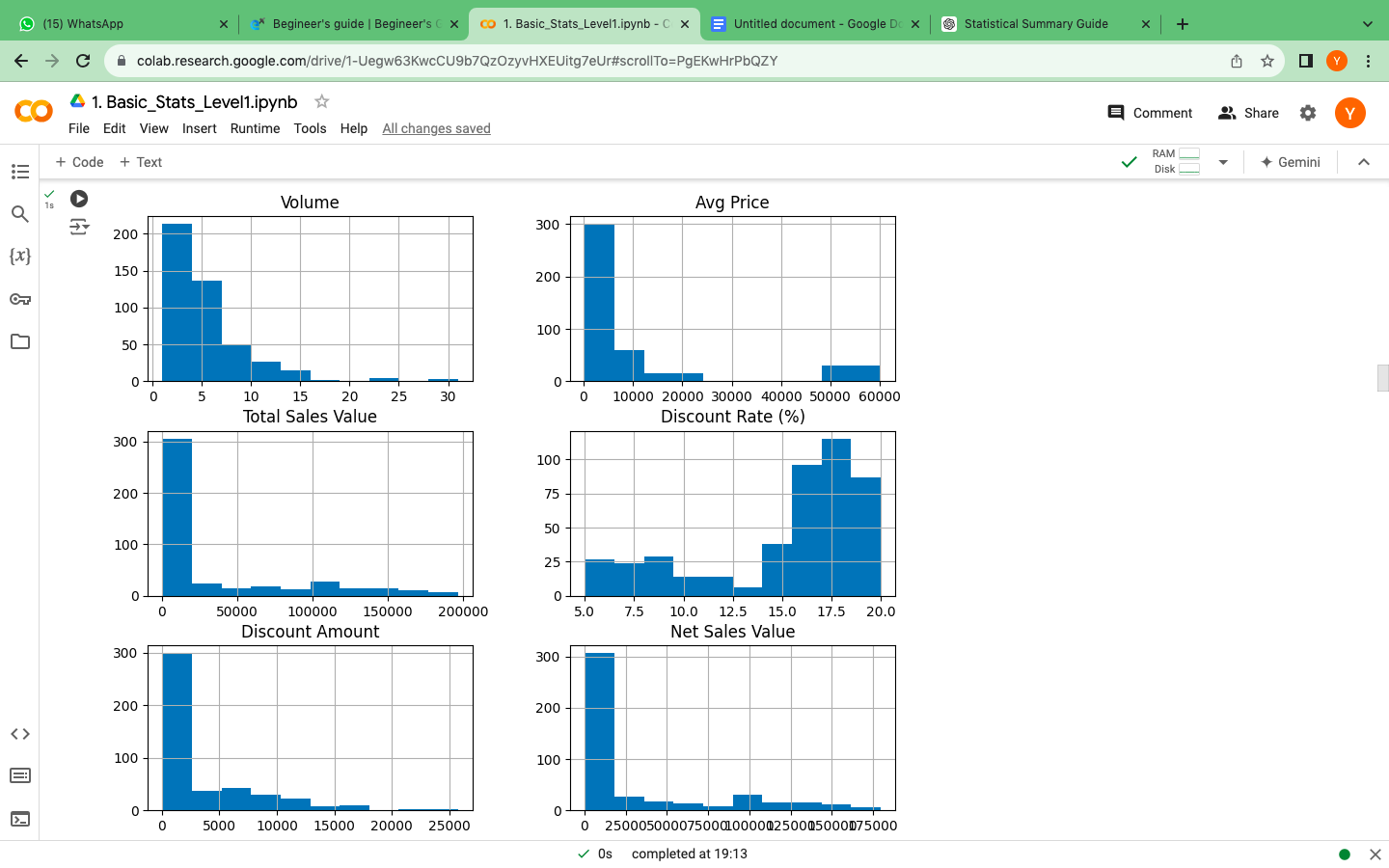
\* Discount Amount

\* Net Sales Value



***HISTOGRAM***

Now let's check the distribution of each numerical column:



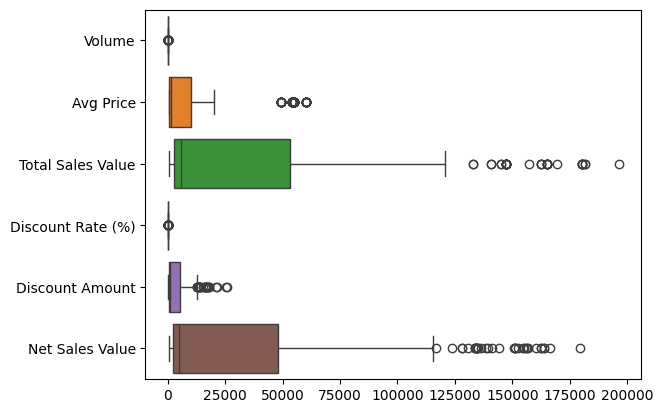
Columns which are somewhat normally distributed:

* Total sales value
* Net sales value

Columns which are negative skewed:

* Discount rate (%)

***BOXPLOT***

******

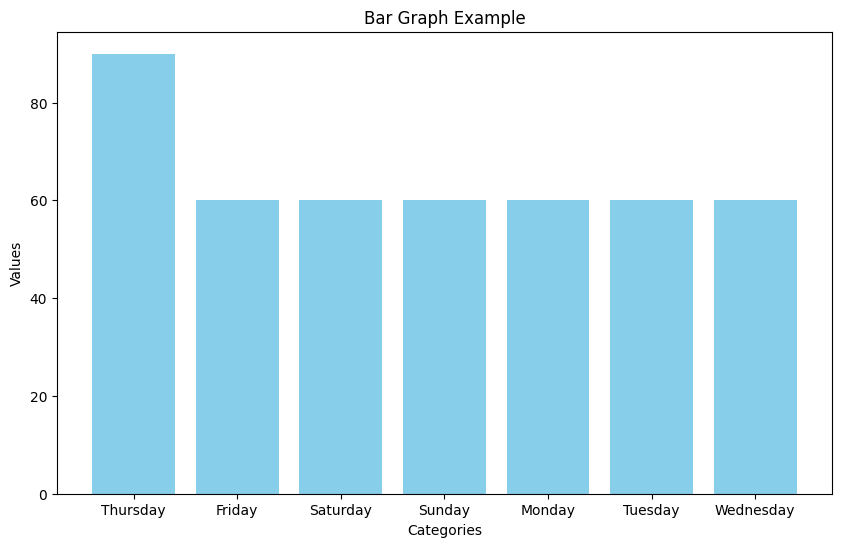
Columns which have potential outliers:

* Total sales value
* Net sales value
* Discount amount

***BARCHART FOR ANALYSIS OF CATEGORICAL COLUMNS***

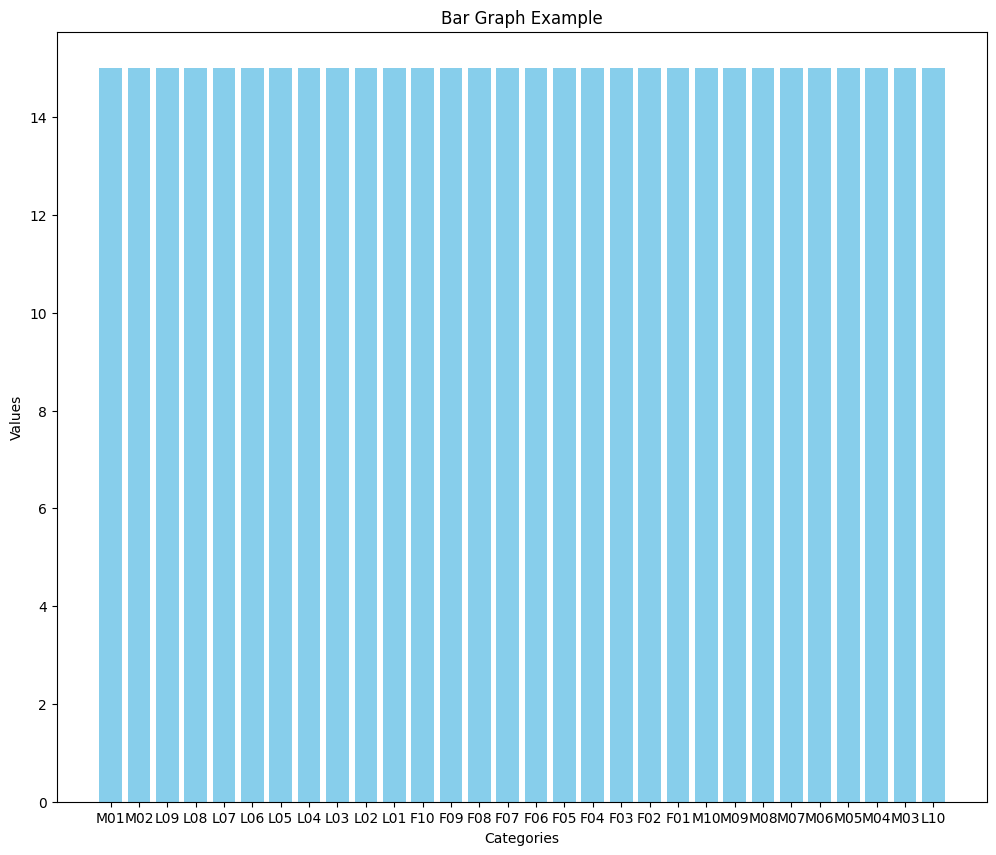
In ‘Day’ column:

* There are 7 categories.
* ‘Thursday’ has more values, and the rest of the categories has same values.

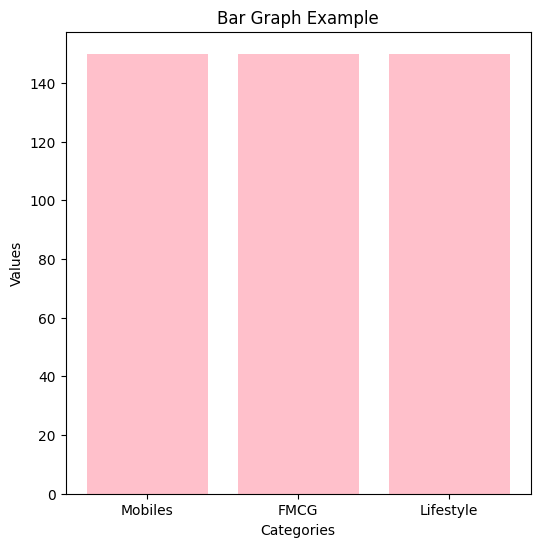
******

In ‘SKU’ column:

* There are 30 categories.
* All of the categories have the same number of value counts..

******

In ‘BU’ column:

* There are 3 categories.
* All of the categories have the same number of value counts..
* Standardisation Of Numerical Variables

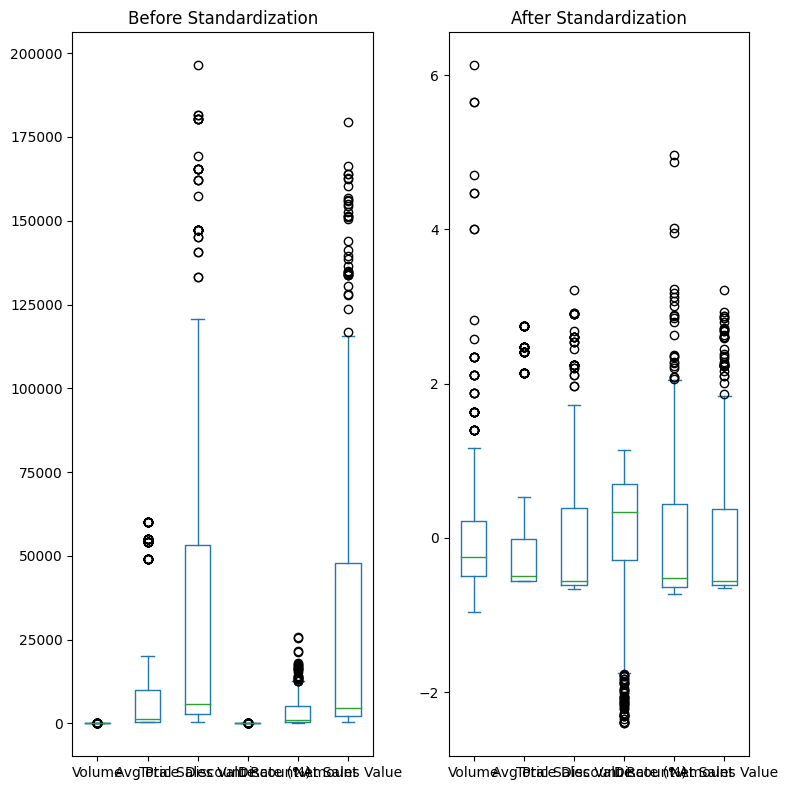
Standardisation is a technique used to transform data to have a mean of 0 and a standard deviation of 1.

This transformation is helpful when comparing features that are on different scales or when using algorithms that assume data is normally distributed.

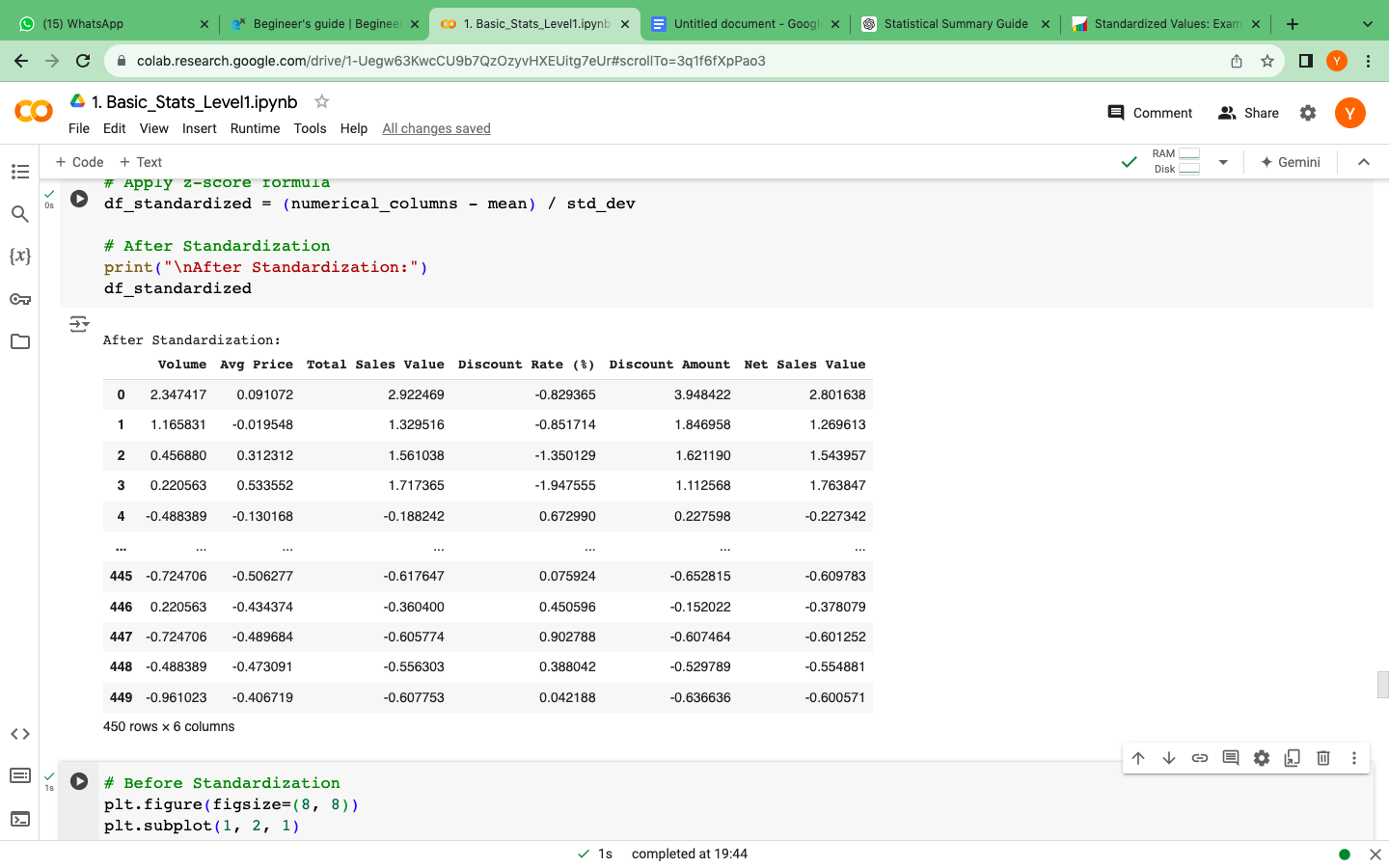
The formula for standardisation is:



where:

* x is the original value,
* μ is the mean of the original data,
* σ is the standard deviation of the original data.

Values after standardisation:

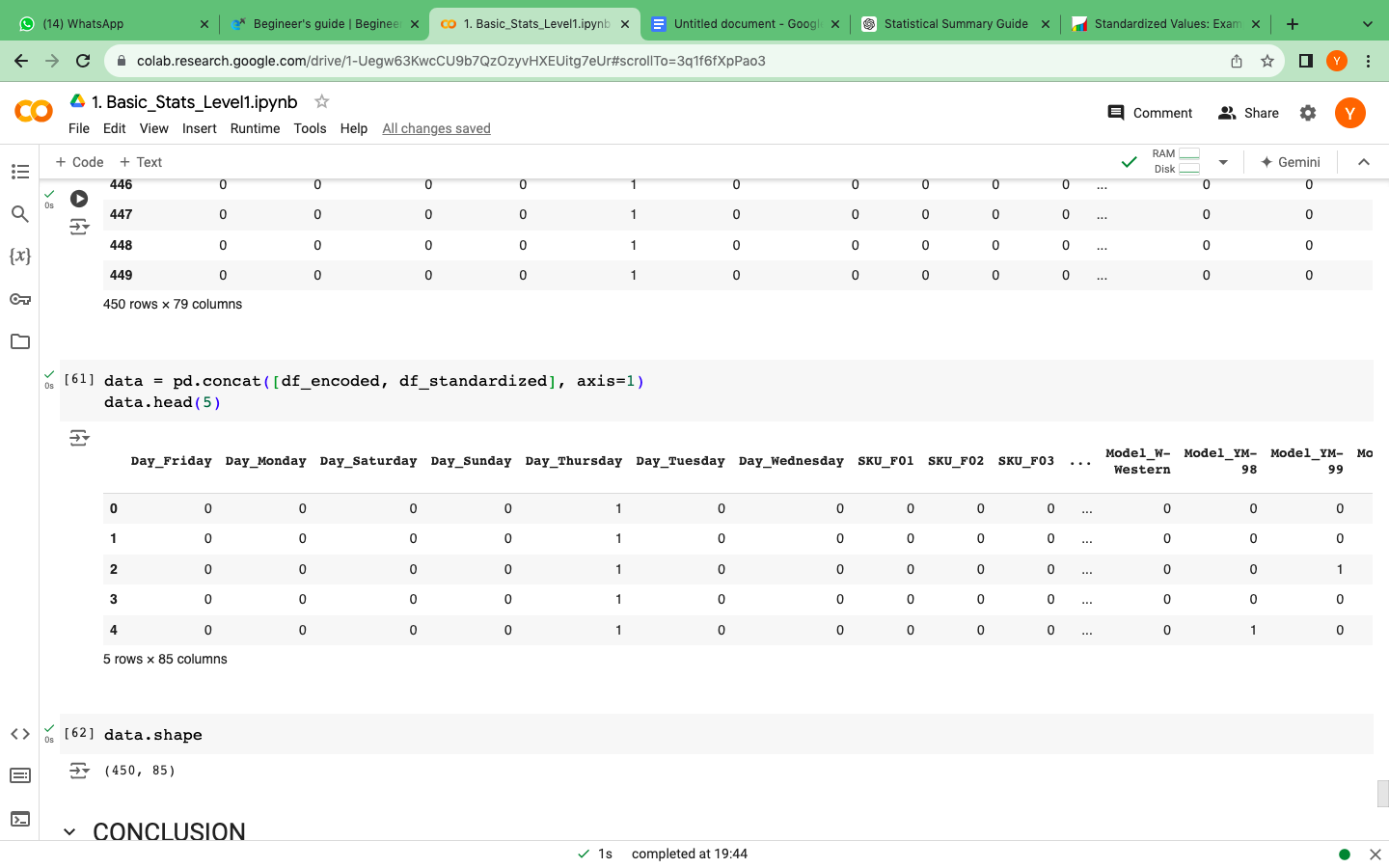


* Conversion Of Categorical Variables Into Dummies

In some machine learning algorithms, categorical data needs to be converted into a number because these algorithms require numerical input.

One method for this is **one-hot encoding**, it transforms categorical variables into binary numbers(0 or 1)

One-hot encoding creates a sparse matrix where each category is represented by a binary vector, making it easy for algorithms to process the data.

***Data after one-hot encoding:***

* Conclusion

1. The shape of data has changed,

* BEFORE ~ (450,10)
* AFTER ~ (450, 85

1. Our data has no missing values.
2. Some columns have outliers.
3. There is skewness in data, i.e the data is not normally distributed.

\* Preprocessing steps like standardisation and one-hot encoding

improve the overall quality of the data.

\* Preprocessing helps in reducing bias and variance in the model by ensuring that all features are appropriately scaled and represented.

\* Preprocessing makes models more robust to variations in the data.

\* Preprocessed data, especially with techniques like one-hot encoding, provides transparency in how categorical variables are handled, aiding in model interpretability and trustworthiness.