What are “Excel Datasets”?

Excel datasets are collections of data that are stored and organized in an Excel spreadsheet, which is a commonly used software that enables users to create, manipulate and analyze data in a structured format. These datasets can come in two main formats: Excel(.xlsx) and Comma Separated Values (CSV). The Excel format provides more advanced features for organizing and analyzing complex data, including the use of formulas and visualizations, while CSV, on the other hand, offers a simpler format that is compatible with a wide range of software applications, making it easier to share data between different programs.

In this article, we have compiled a list of *15 Excel Datasets for Data Analytics Beginners*. With these Excel datasets covering topics like financial analysis, market analysis and time series analysis, beginners can practice data analysis techniques such as data cleaning, pivot tables and charts while gaining insights into real-world scenarios.

List of the Excel Datasets for Data Analytics Beginners

1. Superstore Sales
2. Iris
3. Titanic
4. Wine Quality
5. Adult Census Income
6. Boston Housing
7. Breast Cancer Wisconsin Dataset
8. Online Shoppers Purchasing Intention
9. Bank Marketing
10. Avocado Prices
11. Amazon Top 50 Bestselling Books 2009 – 2019
12. FIFA World Cup
13. New York City Airbnb Open Data
14. World Happiness Report
15. Stock Price

1. Superstore Sales

The Superstore Sales data provides sales data for a fictional retail company, including information on products, orders and customers. It is often used to practice data analytics.

This Excel dataset includes the following variables:

- Order ID - A unique identifier for each order.
- Customer ID - A unique identifier for each customer.
- Order Date - The date of the order placement.
- Ship Date - The date the order was shipped.
- Ship Mode - The shipping mode for the order (e.g. standard, same-day).
- Segment - The customer segment (e.g. Consumer, Corporate, Home Office).
- Region - The region where the customer is located (e.g. West, Central, East).
- Category - The category of the product purchased (e.g. Furniture, Technology, Office Supplies).
- Sub-Category - The sub-category of the product purchased (e.g. Chairs, Desktops, Paper).
- Product Name - The name of the product purchased.
- Sales - The sales revenue for the product purchased.
- Quantity - The number of units of the product purchased.
- Discount - The discount applied to the product purchased.
- Profit -The profit generated by the product purchased.

2. Iris

This dataset includes measurements of the sepal length, sepal width, petal length and petal width of 150 iris flowers, which belong to 3 different species: setosa, versicolor and virginica. The iris dataset has 150 rows and 5 columns, which are stored as a dataframe, including a column for the species of each flower.

The description of its variables includes:

- Sepal.Length - The sepal.length represents the length of the sepal in centimetres.
- Sepal.Width - The sepal.width represents the width of the sepal in centimetres.
- Petal.Length - The petal.length represents the length of the petal in centimetres.
- Species - The species variable represents the species of the iris flower, with three possible values: setosa, versicolor and virginica.

One use case of the Iris dataset in Excel is to analyze the relationship between the different features of the Iris flower and classify the flower species based on the feature values. This can be done using techniques such as correlation analysis, inferential statistics, and predictive modeling.

You can also download this Excel dataset on Kaggle by clicking here.

3. Titanic

This popular open-source dataset offers information on the passengers onboard the Titanic ship when it sank on April 15, 1912. It can be used by data analytics beginners interested in data cleaning and preprocessing, descriptive statistics, data visualization and predictive modeling.

Some of the variables included in the dataset:

- PassengerId - A unique identifier for each passenger.
- Survived - This shows whether the passenger survived or not (0 = No, 1 = Yes).
- Pclass - A passenger's class (1 = 1st, 2 = 2nd, 3 = 3rd).
- Name - A passenger's name.
- Sex - A passenger's gender.
- Age - A passenger's age.
- SibSp - The number of siblings/spouses aboard.
- Parch - The number of parents/children aboard.
- Ticket - The ticket number.
- Fare - The fare paid for the ticket.
- Cabin - The cabin number.
- Embarked - The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

4. Wine Quality

The Wine Quality dataset contains information on red and white wine samples. This dataset aims to classify the quality of the wine based on chemical properties like pH, density, alcohol content and citric acid content.

The common variables included in this Excel dataset:

- Fixed Acidity - The number of fixed acids in the wine, expressed in g/dm^3.
- Volatile Acidity - The number of volatile acids in the wine, expressed in g/dm^3.

- Citric Acid - The amount of citric acid in the wine, expressed in g/dm^3.
- Residual Sugar - The amount of residual sugar in the wine, expressed in g/dm^3
- Chlorides - The amount of chloride in the wine, expressed in g/dm^3.
- Free Sulfur Dioxide - The amount of free sulfur dioxide in the wine, expressed in mg/dm^3.
- Total Sulfur Dioxide - The amount of total sulfur dioxide in the wine, expressed in mg/dm^3.
- Density - The density of the wine, expressed in g/cm^3.
- pH - The pH level of the wine.
- Sulphates - The number of sulphates in the wine, expressed in g/dm^3.
- Alcohol - The alcohol content of the wine, expressed in % vol.
- Quality - The quality rating of the wine, on a scale of 0 to 10.

5. Adult Census Income

This Excel dataset is a collection of information about individuals living in the United States, extracted from the 1994 Census database. It contains various demographic, social and economic attributes about each individual.

Some of the attributes included in this dataset:

- age
- Workclass - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt
- Education - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- Education-num
- marital-status - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

- occupation - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex - Male or female.

The “income” attribute is the target variable and the dataset is very useful to data analytics beginners.

6. Boston Housing

The Boston Housing dataset consists of information on housing in the area of Boston, Massachusetts. It has about 506 rows and 14 columns of data.

Some of the variables in the dataset include:

- CRIM - Per capita crime rate by town.
- ZN - The proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - The proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- NOX - The nitric oxide concentration (parts per 10 million).
- RM - The average number of rooms per dwelling.
- AGE - The proportion of owner-occupied units built prior to 1940.

- DIS - The weighted distances to five Boston employment centres.
- RAD - The Index of accessibility to radial highways.
- TAX - The full-value property-tax rate per $10,000.
- PTRATIO - The pupil-teacher ratio by town.
- B - 1000(Bk - 0.63)^2 where -Bk is the proportion of blacks by town.
- LSTAT - The percentage lower status of the population.
- MEDV - The median value of owner-occupied homes in $1000's.

This dataset can be utilized in data analytics to analyze the relationship between various features of house prices and a housing market, perform data analysis and generate insights.

7. Breast Cancer Wisconsin Dataset

This Excel dataset consists of information about breast cancer tumours and was initially created by Dr. William H. Wolberg. The dataset was created to assist researchers and machine learning practitioners in classifying tumours as either malignant(cancerous) or benign (non-cancerous).

Some of the variables included in this dataset:

- ID number
- Diagnosis (M = malignant, B = benign).
- Radius (the mean of distances from the centre to points on the perimeter).
- Texture (the standard deviation of gray-scale values).
- Perimeter
- Area
- Smoothness (the local variation in radius lengths).
- Compactness (the perimeter^2 / area - 1.0).
- Concavity (the severity of concave portions of the contour).
- Concave points (the number of concave portions of the contour).

- Symmetry
- Fractal dimension ("coastline approximation" - 1).

8. **Online Shoppers Purchasing Intention**

The Online Shoppers Purchasing Intention dataset is a collection of data related to purchase patterns and consumer behaviour in the context of online shopping. It was created by conducting surveys of online shoppers and collecting data from their responses.

Some of the variables in this dataset include:

- Administrative - The number of pages of the website visited by the user for administrative purposes
- Administrative_Duration - The total time spent by the user on administrative pages of the website
- Informational - The number of pages of the website visited by the user for informational purposes
- Informational_Duration - The total time spent by the user on informational pages of the website
- ProductRelated - The number of pages of the website visited by the user for product-related purposes
- ProductRelated_Duration - The total time spent by the user on product-related pages of the website
- BounceRates - The percentage of visitors who enter the website and leave without viewing any other pages
- ExitRates - The percentage of visitors who exit the website from a particular page after visiting it
- PageValues - The average value of the pages viewed by the user before the transaction
- SpecialDay - The proximity of the visit to a special day (e.g., Mother's Day, Valentine's Day, etc.)

This Excel dataset is used in research and analytics related to e-commerce and online marketing. It can help businesses to understand the factors that drive customer behaviour and is also useful for data analytics beginners.

9. Bank Marketing

This popular dataset is to study marketing campaigns for a Portuguese banking institution. It contains information about the bank's marketing campaigns, as well as customer demographics and economic indicators.

Some of the variables included in this dataset:

- Age - Age of the customer (numeric)
- Job - Type of job
- Marital - Marital status
- Education - Education level
- Default - Has credit in default?
- Balance - Average yearly balance, in euros.
- Housing - Has a housing loan?
- Loan - Has a personal loan?
- Contact - Contact communication type.
- Day - Day of the month contacted.
- The output variable denotes whether or not the customer subscribed to a term deposit after being contacted by the bank.

10. Avocado Prices

The Avocado Prices dataset consists of data related to the prices of avocados in the United States. The data is collected from various sources like the Hass Avocado Board and the United States Department of Agriculture (USDA).

Some of the variables in this dataset include:

- Date - The date of the observation.

- AveragePrice - The average price of a single avocado.
- Total Volume - Total number of avocados sold.
- PLU (Price Look-Up) code - A code used to identify a specific type of avocado.
- Type - Conventional or organic
- Region - The city or region of the observation.

It can also be used by businesses in the food industry to make strategic decisions about buying and selling avocados.

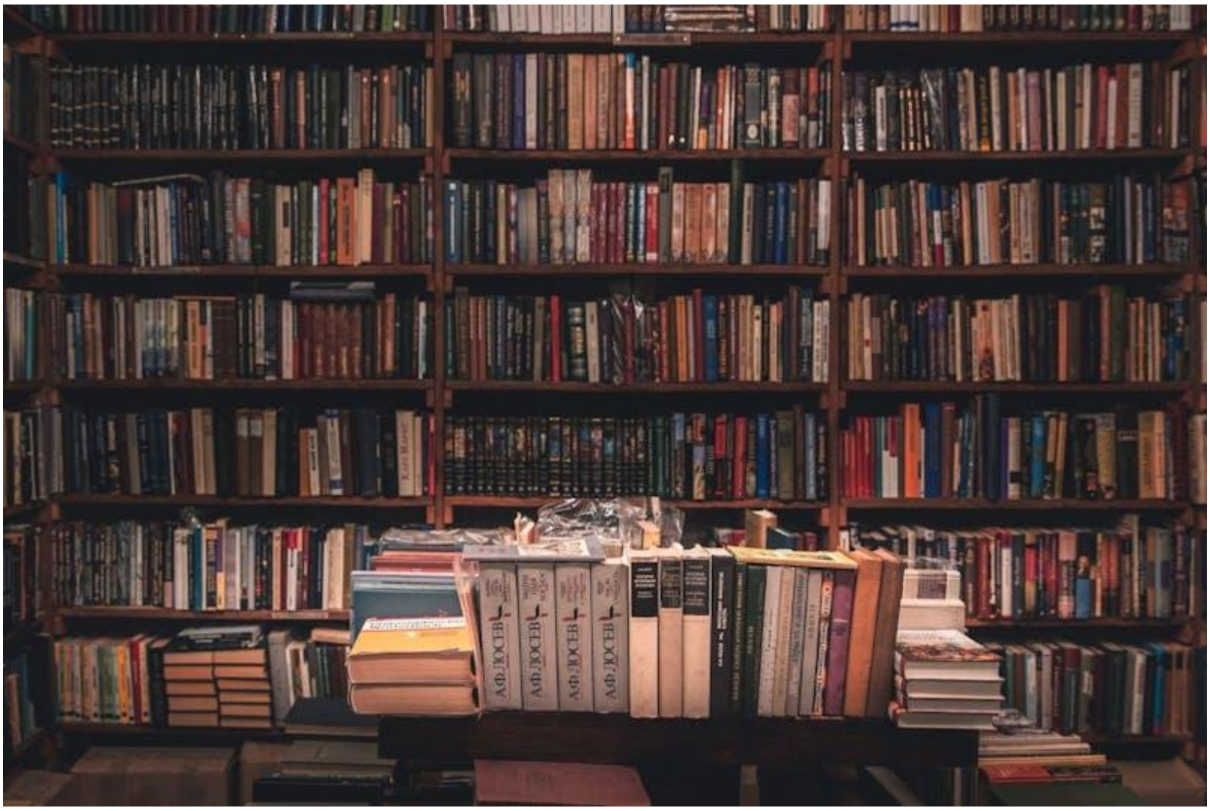
11. Amazon Top 50 Bestselling Books 2009 - 2019

This Excel dataset is a collection of data related to the top 50 best-selling books on Amazon for each year between 2009 and 2019.

The dataset includes the following variables:

- Name - The title of the book.
- Author - The name of the book's author.
- User Rating - The average rating of the book as provided by Amazon users.
- Reviews - The total number of reviews the book has received on Amazon.
- Price - The price of the book in US dollars.
- Year - The year the book was published.
- Genre - The genre of the book.

The Amazon Top 50 Bestselling Books can be used to explore trends in book sales on Amazon over a decade and is useful to data analytics beginners.



12. FIFA World Cup

The FIFA World Cup dataset is a collection of data related to the FIFA World Cup which is held every four years. It contains information on every World Cup tournament from 1930 to 2014.

Some of the variables in this dataset include:

- Year - The year of the tournament.
- Country - The host country of the tournament.
- Winner - The team that won the tournament.
- Runners-Up - The team that finished as the runners-up.
- Third - The team that finished in third place.
- Fourth - The team that finished in fourth place.
- GoalsScored - The total number of goals scored in the tournament.
- QualifiedTeams - The total number of teams that qualified for the tournament.

- Attendance - The total number of spectators who attended the matches.

The dataset can be used to analyze trends in the World Cup over time, such as changes in the number of teams that participate or the number of goals scored.

13. New York City Airbnb Open Data

This excel dataset consists of public information about Airbnb listings and metrics in New York City. The 2019 New York City Airbnb Open Data includes information on about 50,000 Airbnb listings in the city and is made available to the public by the New York City government to promote transparency and understanding of the impact of rentals on the city.

Some of the variables in the dataset include:

- Id - A unique identifier for each Airbnb listing.
- Name - The name of the Airbnb listing.
- Host_id - A unique identifier for the Airbnb host.
- Host_name - The name of the Airbnb host.
- Neighbourhood_group - The borough of the Airbnb listing.
- Neighbourhood - The neighbourhood of the Airbnb listing.
- Latitude - The latitude of the Airbnb listing.
- Longitude - The longitude of the Airbnb listing.
- Room_type - The type of room available for rent (e.g. private room, entire home/apt, shared room).
- Price - The nightly price to rent the Airbnb listing.

14. World Happiness Report

This dataset includes information on the happiness levels of over 150 countries, such as economic, social, and health factors that contribute to happiness. It is useful to data analytics beginners for practicing data exploration, visualization, and regression analysis.

Some of the variables in this dataset include:

- Country name - Name of the country.
- Year - Year of the survey.
- Life Ladder - Average life satisfaction score based on a scale of 0-10.
- Log GDP per capita - Natural logarithm of GDP per capita, adjusted for purchasing power parity (PPP) in constant 2017 international dollars.
- Healthy life expectancy at birth - The expected number of years to be lived in full health, adjusted for years spent in poor health.

15. Stock Price

This dataset includes the daily stock prices of various companies, such as Apple, Google and Amazon. It is useful for practicing time series analysis and predicting future stock prices.

The variables in this dataset:

- Date - The date when the stock price was recorded.
- Open - The opening price of the stock.
- High - The highest price of the stock during the trading day.
- Low - The lowest price of the stock during the trading day.
- Close - The closing price of the stock.
- Adj Close - The adjusted closing price of the stock.
- Volume - The number of shares traded during the day.

Common Practice Questions for These Excel Datasets

Superstore Sales

- What is the total revenue generated by the store?
- Which category of products contributes the most to sales?
- How has the sales trend been for the past year?
- Which region has the highest sales and which one has the lowest?
- What is the average profit margin of the store?

Iris

- What is the distribution of each species of iris in the dataset?
- What is the correlation between petal length and petal width?
- What is the average sepal length for each species of iris?
- Which species of iris has the largest petal area?
- How many observations are there for each species of iris?

Titanic

- What is the survival rate of the passengers?
- What is the average age of the passengers?
- What is the proportion of male and female passengers?
- Which class of passengers had the highest survival rate?
- What is the distribution of the fare paid by the passengers?

Wine Quality

- What is the correlation between pH and alcohol content?
- Which type of wine (red or white) has a higher median quality rating?
- What is the median volatile acidity for each type of wine?
- What is the proportion of each wine type in the dataset?
- What is the distribution of citric acid for each wine type?

Adult Census Income

- What is the proportion of people who earn more than $50K?
- What is the average age of people who earn more than $50K?
- What is the correlation between age and education level?
- What is the proportion of men and women who earn more than $50K?
- What is the median hours worked per week for people who earn more than $50K?

Boston Housing

- What is the correlation between the number of rooms and the median value of owner-occupied homes?
- Which variable has the highest correlation with the median value of owner-occupied homes?
- What is the average age of the homes?
- What is the distribution of the pupil-teacher ratio by town?
- Which town has the highest median value of owner-occupied homes?

Breast Cancer Wisconsin Dataset

- What is the proportion of benign and malignant tumours?
- What is the correlation between tumour radius and perimeter?
- What is the average smoothness of the tumours?

- What is the distribution of the concavity of the tumours?
- What is the median area of the tumours?

Online Shoppers Purchasing Intention

- What is the proportion of visitors who made a purchase?
- What is the distribution of the number of pages visited by the visitors?
- What is the average time spent on the website by the visitors?
- What is the correlation between the bounce rate and the revenue?
- What is the distribution of the operating system used by the visitors?

Bank Marketing

- What is the proportion of people who subscribed to a term deposit?
- What is the correlation between age and balance?
- What is the distribution of the job type of the customers?
- What is the average duration of the calls?
- What is the proportion of calls made each month?

Amazon Top 50 Bestselling Books 2009 – 2019

- What is the average rating of the books?
- What is the distribution of the number of reviews received by the books?
- Which book has the highest price?
- What is the correlation between the rating and the price of the books?
- What is the distribution of the genres of the books?

FIFA World Cup

- What is the average number of goals scored per game?
- What is the proportion of games that ended in a draw?
- Which country has won the most World Cup titles?
- What is the average age of players in the tournament?
- What is the distribution of attendance for each game?

New York City Airbnb Open Data

- What is the average price of the listings?

- What is the distribution of the room types available for the listings?
- Which neighbourhood has the most listings?
- What is the correlation between the number of reviews and the price of the listings?
- What is the distribution of the cancellation policies for the listings?

World Happiness Report

- What is the distribution of the happiness scores for each country?
- Which country has the highest happiness score?
- What is the correlation between GDP per capita and happiness score?
- What is the distribution of the factors that contribute to happiness?
- Which region of the world has the highest average happiness score?

Stock Price

- What is the average daily return of the stock?
- What is the distribution of the daily trading volume? Avocado Prices
- What is the average price of avocados?
- What is the distribution of the average price by region?
- Which region has the highest and lowest average price?
- What is the correlation between the total volume and the average price?
- What is the distribution of the total volume by year?