Business Statistics Course Datasets

This document lists recommended datasets for exploratory data analysis (EDA), descriptive statistics, and inferential analysis in business contexts.
Each dataset includes a short description, purpose, and a link to the source.

1. Overview

These datasets are chosen to cover a range of business applications — from marketing and HR to finance and sustainability — and can be used for:

- Data cleaning and preparation
- Exploratory and descriptive analysis
- Hypothesis testing and inferential statistics
- Regression and predictive modeling

2. Recommended Datasets

Purpose	Dataset	Why It's Good
General business data	Superstore Sales	Rich, balanced, and intuitive dataset for analyzing sales, profit, and regional performance. Excellent for descriptive stats and regression.
Marketing analysis	Bank Marketing	Perfect for classification and inference tasks; includes demographics, marketing methods, and campaign results. Great for logistic regression and hypothesis testing.
HR analytics	IBM Attrition	People-focused dataset; ideal for exploring employee turnover, job satisfaction, and salary trends using descriptive and inferential methods.
Time series & finance	S&P 500 / Stock Data	Great for time series analysis, regression, and volatility exploration using real market data.
E-commerce	Online Retail (UCI)	Real-world transactions for segmentation and sales trend analysis; ideal for customer behavior insights.
Shared economy / real estate	Airbnb Listings (Various Cities)	Combines numeric and categorical variables—excellent for EDA, regression, and hypothesis testing (e.g., does room type affect price or rating?).
Sustainability / ESG	World Sustainability Dataset	Longitudinal and cross-sectional data on sustainability indicators across 173 countries; perfect for descriptive and comparative analysis.

3. Suggested Analyses by Dataset

This section outlines the types of exploratory, descriptive, and inferential analyses that can be performed with each dataset.

3.1 Superstore Sales Dataset

Exploratory / Descriptive:

- Examine sales and profit distributions by category, region, and customer segment
- Compute summary statistics (mean, median, variance) for sales and discounts
- Create correlation heatmaps between numeric variables (e.g., sales, discount, profit)
- Visualize sales trends over time

Inferential / Predictive:

- Hypothesis testing: Does offering a discount significantly affect profit margins?
- ANOVA: Compare mean profit across product categories
- Regression: Predict profit using sales, discount, and shipping cost
- Time series decomposition of monthly sales for trend and seasonality

---

3.2 Bank Marketing Dataset

Exploratory / Descriptive:

- Analyze customer demographics (age, job type, education, marital status)
- Examine campaign outcomes across contact types (phone, cellular, unknown)
- Visualize distributions and correlations among numeric variables (age, balance, duration)

Inferential / Predictive:

- Chi-square tests: Relationship between contact type and campaign success
- Logistic regression: Predict probability of customer subscribing to a term deposit
- Cross-tab analysis by education and response rate
- Evaluate classification metrics (accuracy, precision, recall)

3.3 IBM HR Attrition Dataset

Exploratory / Descriptive:

- Profile employees by department, age, salary, and satisfaction level
- Compare attrition rates across job roles and performance ratings
- Visualize categorical relationships (department vs. attrition, gender vs. attrition)

Inferential / Predictive:

- Chi-square tests for categorical relationships (attrition vs. gender, marital status)
- t-tests for mean salary differences between those who left and those who stayed
- Logistic regression to predict employee attrition based on tenure, satisfaction, and salary
- Correlation and regression analysis between performance metrics and compensation

---

3.4 S&P 500 / Stock Data

Exploratory / Descriptive:

- Examine daily returns, volatility, and volume trends
- Plot moving averages, correlations among sectors, and overall market index behavior
- Identify outliers and anomalies in stock returns

Inferential / Predictive:

- Time series regression: Model closing price as a function of lagged returns
- Hypothesis testing: Are mean returns significantly different across sectors?
- ARIMA or exponential smoothing for forecasting future prices
- Correlation analysis between stock returns and macroeconomic indicators (if available)

3.5 Online Retail (UCI)

Exploratory / Descriptive:

- Analyze customer purchase frequency, average basket size, and total revenue
- Segment customers based on RFM (Recency, Frequency, Monetary value)
- Identify top-selling products and seasonal trends

Inferential / Predictive:

- Hypothesis testing: Do customers from different countries spend differently?
- Regression: Predict total spending per customer
- ANOVA for comparing sales across product categories
- Cohort analysis to study customer retention over time

3.6 Airbnb Listings Dataset

Exploratory / Descriptive:

- Summarize price, number of reviews, and availability by room type and location
- Explore correlations between price, rating, and amenities
- Map geographic distribution of listings and pricing

Inferential / Predictive:

- Hypothesis testing: Does room type affect average rating?
- Regression: Predict price using features such as location, number of reviews, and property type
- ANOVA for comparing prices across neighborhoods
- Cluster analysis for identifying market segments

3.7 World Sustainability Dataset

Exploratory / Descriptive:

- Summarize sustainability indicators by country, region, and year
- Visualize trends in emissions, renewable energy use, or education levels
- Correlate GDP per capita with environmental and social indicators

Inferential / Predictive:

- Hypothesis testing: Do developed and developing countries differ in sustainability performance?
- Regression: Predict sustainability index using GDP, population, and education variables
- Time series or panel data analysis for long-term trends
- Cluster analysis to identify groups of countries with similar sustainability profiles

---

4. Suggested Course Flow (Example Using Superstore Sales)

Week	Topic	Activity
1	Data Cleaning & Overview	Load CSV, inspect variables, handle missing data
2	Descriptive Statistics	Compute mean, median, mode, and dispersion by category
3	Exploratory Visualization	Create histograms, boxplots, scatterplots, and heatmaps
4	Hypothesis Testing	Test if discount rate affects average profit
5	Regression Analysis	Predict profit using sales, discount, and category
6	Reporting Insights	Develop a short business report with key findings and visuals

---

5. Notes

- All datasets are free to download and can be used in **Excel, R, Python, or Power BI**.
- For smaller workshops or short courses, start with **Superstore Sales** or **HR Attrition**.
- For advanced or MBA-level sessions, integrate **S&P 500** or **Airbnb** for richer analysis.
- Encourage students to explore relationships between variables using **both visualization and statistical inference**.
