

Projet Python - Pandas

Objectifs

Analyser l'impact de la crise sanitaire sur le marché immobilier français, par région et département.

Analyser l'impact de la hausse des taux d'intérêt sur le marché immobilier français, par région et département.

Identifier les variables susceptibles d'être utilisées dans la prévision des prix de l'immobilier.

Réaliser un modèle prédictif des prix de l'immobilier et en analyser la performance (optionnel / bonus).

Challenges du projet

- Volume : données relativement volumineuses
- Structure : données de panel (ou données longitudinales) des données qui comprennent plusieurs observations au cours du temps pour un même individu (ville dans notre cas)

Jeu de données

**RÉPUBLIQUE
FRANÇAISE**
Liberté
Égalité
Fraternité

data.gouv.fr

Se connecter S'enregistrer

Recherche

Données Réutilisations Organisations Commencer sur data.gouv.fr Actualités Nous contacter

Accueil > Jeux de données > DVF

Ajouter aux favoris

Demandes de valeurs foncières DVF

Description

Propos liminaires

Conformément au décret n° 2018-1350 du 28 décembre 2018 relatif à la publication sous forme électronique des informations portant sur les valeurs foncières déclarées à l'occasion des mutations immobilières, le présent fichier DVF est désormais disponible en open data.

La publication de ces données répond à l'objectif de transparence des marchés fonciers et immobiliers.

Le fichier DVF contient des données à caractère personnel et la DGFIP attire votre attention sur les obligations légales qui en découlent :

Producteur

[Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique](#)

Dernière mise à jour

10 octobre 2023













Licence

[Licence Ouverte / Open Licence version 2.0](#)

 Qualité des métadonnées

<https://www.data.gouv.fr/fr/datasets/5c4ae55a634f4117716d5656/>

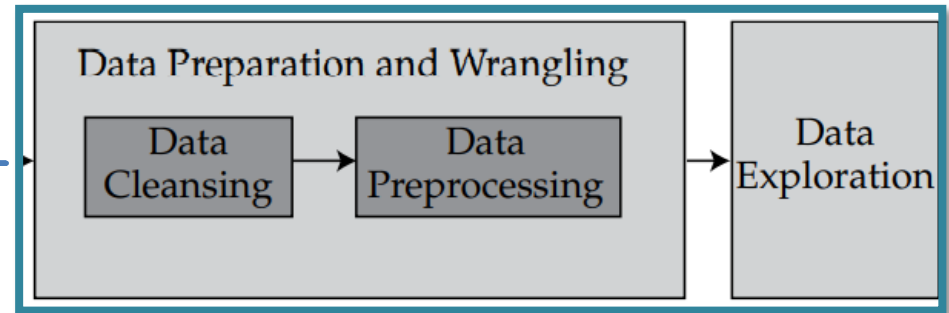
Jeu de données

 Valeurs foncières 2023 - Premier semestre Mis à jour le 10 octobre 2023 — txt (195.2Mo) — 22556 téléchargements	▼ Voir les métadonnées	
 Valeurs foncières 2022 Mis à jour le 10 octobre 2023 — txt (590.7Mo) — 47216 téléchargements	▼ Voir les métadonnées	
 Valeurs foncières 2021 Mis à jour le 10 octobre 2023 — txt (596.7Mo) — 17743 téléchargements	▼ Voir les métadonnées	
 Valeurs foncières 2020 Mis à jour le 10 octobre 2023 — txt (448.9Mo) — 6964 téléchargements	▼ Voir les métadonnées	
 Valeurs foncières 2019 Mis à jour le 10 octobre 2023 — txt (462.8Mo) — 11555 téléchargements	▼ Voir les métadonnées	
 Valeurs foncières 2018 - Second semestre Mis à jour le 10 octobre 2023 — txt (231.0Mo) — 7928 téléchargements	▼ Voir les métadonnées	

Etapes du projet

- Profilage des données
https://fr.wikipedia.org/wiki/Data_profiling

- Nettoyage
- Préparation
- Analyse exploratoire
- Réponses aux objectifs



Livrable

1 fichier zip contenant

- 1 Notebook Jupyter
 - avec les titres et sous-titres des parties
 - un code clairement commenté et lisible
- 1 rapport final d'analyse d'au moins 15 pages contenant
 - une présentation du sujet ;
 - une présentation des différentes étapes du projet
 - les constatations et choix faits au cours de ces étapes
 - les traitements effectués
 - les difficultés rencontrées et les solutions appliquées le cas échéant
 - les analyses et conclusions relatives aux objectifs énoncés (illustrés notamment par des graphiques)
 - une conclusion globale contenant une ouverture

Bases théoriques pour la réalisation du projet

Steps in executing a data analysis project

Structured data

- 1 Conceptualization of the modeling task.** This crucial first step entails determining what the output of the model should be (e.g., whether the price of a stock will go up/down one week from now), how this model will be used and by whom, and how it will be embedded in existing or new business processes
- 2 Data collection.** The data traditionally used for financial forecasting tasks are mostly numeric data derived from internal and external sources. Such data are typically already in a structured tabular format, with columns of features, rows of instances, and each cell representing a particular value.
- 3 Data preparation and wrangling.** This step involves cleansing and preprocessing of the raw data. Cleansing may entail resolving missing values, out-of-range values, and the like. Preprocessing may involve extracting, aggregating, filtering, and selecting relevant data columns.
- 4 Data exploration.** This step encompasses exploratory data analysis, feature selection, and feature engineering.
- 5 Model training.** This step involves selecting the appropriate ML method (or methods), evaluating performance of the trained model, and tuning the model accordingly.

Note that **these steps are iterative** because model building is an iterative process. The insights gained from one iteration may inform the next iteration, beginning with reconceptualization.

Steps in executing a data analysis project

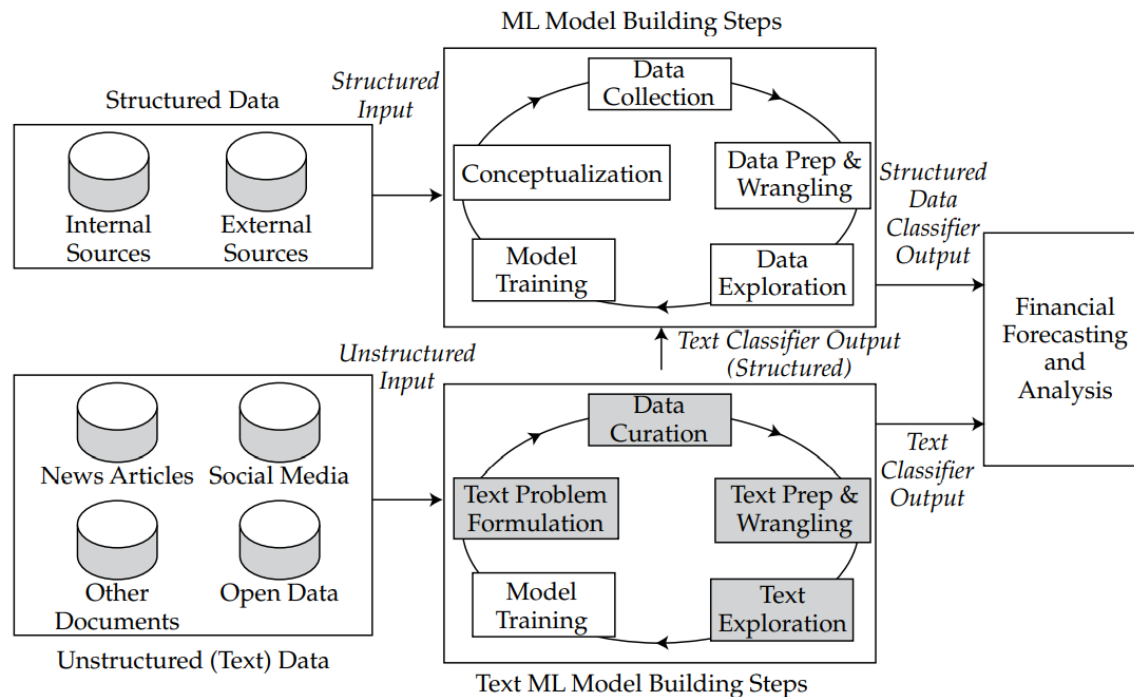
Textual data

The Text ML Model Building **Steps used for the unstructured data sources** of big data are shown in the bottom half of Exhibit 1. They differ from those used for traditional data sources and are typically intended to create output information that is structured.

- 1 Text problem formulation.** Analysts begin by determining how to formulate the text classification problem, identifying the exact inputs and outputs for the model. Perhaps we are interested in computing sentiment scores (structured output) from text (unstructured input). Analysts must also decide how the text ML model's classification output will be utilized.
- 2 Data (text) curation.** This step involves gathering relevant external text data via web services or web spidering (scraping or crawling) programs that extract raw content from a source, typically web pages. Annotation of the text data with high-quality, reliable target (dependent) variable labels might also be necessary for supervised learning and performance evaluation purposes. For instance, experts might need to label whether a given expert assessment of a stock is bearish or bullish.
- 3 Text preparation and wrangling.** This step involves critical cleansing and preprocessing tasks necessary to convert streams of unstructured data into a format that is usable by traditional modeling methods designed for structured inputs.
- 4 Text exploration.** This step encompasses text visualization through techniques, such as word clouds, and text feature selection and engineering.

Steps in executing a data analysis project

Exhibit 1 Model Building for Financial Forecasting Using Big Data: Structured (Traditional) vs. Unstructured (Text)



Data preparation and wrangling

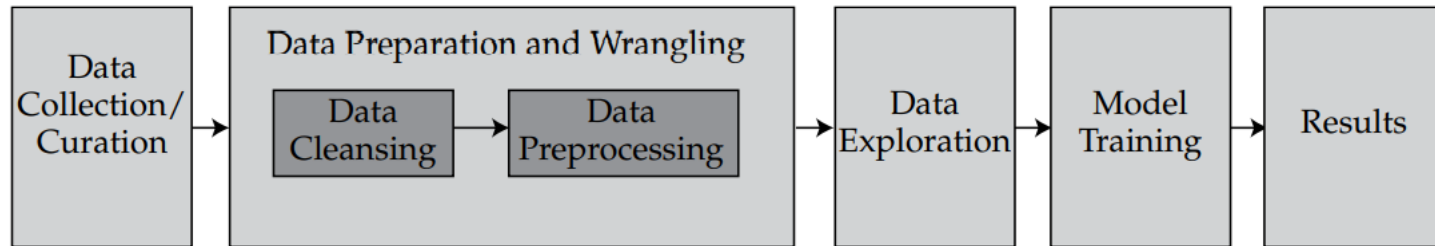
Cleansing and Organizing raw data into a consolidated format. The resulting dataset is suitable to use for further analyses and training a machine learning (ML) model.

This is a **critical stage**, the foundation, in big data projects. **Most of the project time is spent on this step**, and the quality of the data affects the training of the selected ML model.

Domain knowledge—that is, the involvement of specialists in the particular field in which the data are obtained and used—is beneficial and often necessary to successfully execute this step.

Data preparation and wrangling

Exhibit 2 Data Preparation and Wrangling Stage



Data Preparation (Cleansing): This is the initial and most common task in data preparation that is performed on raw data. Data cleansing is the process of examining, identifying, and mitigating errors in raw data. Normally, the raw data are neither sufficiently complete nor sufficiently clean to directly train the ML model. Manually entered data can have incomplete, duplicated, erroneous, or inaccurate values. Automated data (recorded by systems) can have similar problems due to server failures and software bugs.

Data Wrangling (Preprocessing): This task performs transformations and critical processing steps on the cleansed data to make the data ready for ML model training. Raw data most commonly are not present in the appropriate format for model consumption. After the cleansing step, data need to be processed by dealing with outliers, extracting useful variables from existing data points, and scaling the data.

Structured Data Data Preparation (Cleansing)

Structured data are organized in a systematic format that is readily searchable and readable by computer operations for processing and analyzing. In structured data, **data errors can be in the form of incomplete, invalid, inaccurate, inconsistent, nonuniform, and duplicate data observations**. The data cleansing process mainly deals with identifying and mitigating all such errors. **Exhibit 3** shows a raw dataset before cleansing. The data have been collected from different sources and are organized in a data matrix (or data table) format. Each row contains observations of each customer of a US-based bank. Each column represents a variable (or feature) corresponding to each customer.

Exhibit 3 Raw Data Before Cleansing

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	\$50,200	\$5,000	VA	Y
3	2	Ms. XYZ	M	15 Jan, 1975	\$60,500	\$0	NY	Yes
4	3	EFG		1/13/1979	\$65,000	\$1,000	CA	No
5	4	Ms. MNO	F	1/1/1900	—	—	FL	Don't Know

Exhibit 3 (Continued)

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
6	5	Ms. XYZ	F	15/1/1975	\$60,500	\$0	Y	
7	6	Mr. GHI	M	9/10/1942	NA	\$55,000	TX	N
8	7	Mr. TUV	M	2/27/1956	\$300,000	\$50,000	CT	Y
9	8	Ms. DEF	F	4/4/1980	\$55,000	\$0	British Columbia	N

Structured Data Data Preparation (Cleansing)

The possible errors in a raw dataset include the following:

- 1 Incompleteness error is where the data are not present, resulting in missing data.** This can be corrected by investigating alternate data sources. Missing values and NAs (not applicable or not available values) must be either omitted or replaced with “NA” for deletion or substitution with imputed values during the data exploration stage. The most common imputations are mean, median, or mode of the variable or simply assuming zero. In Exhibit 3, rows 4 (ID 3), 5 (ID 4), 6 (ID 5), and 7 (ID 6) are incomplete due to missing values in either Gender, Salary, Other Income, Name (Salutation), and State columns.
- 2 Invalidity error is where the data are outside of a meaningful range, resulting in invalid data.** This can be corrected by verifying other administrative data records. In Exhibit 3, row 5 likely contains invalid data as the date of birth is out of the range of the expected human life span.
- 3 Inaccuracy error is where the data are not a measure of true value.** This can be rectified with the help of business records and administrators. In Exhibit 3, row 5 is inaccurate (it shows “Don’t Know”); in reality, every person either has a credit card or does not.

Structured Data Data Preparation (Cleansing)

The possible errors in a raw dataset include the following:

- 4 Inconsistency error is where the data conflict with the corresponding data points or reality.** This contradiction should be eliminated by clarifying with another source. In Exhibit 3, row 3 (ID 2) is likely to be inconsistent as the Name column contains a female title and the Gender column contains male.
- 5 Non-uniformity error is where the data are not present in an identical format.** This can be resolved by converting the data points into a preferable standard format. In Exhibit 3, the data under the Date of Birth column is present in various formats. The data under the Salary column may also be non-uniform as the monetary units are ambiguous; the dollar symbol can represent US dollar, Canadian dollar, or others.
- 6 Duplication error is where duplicate observations are present.** This can be corrected by removing the duplicate entries. In Exhibit 3, row 6 is a duplicate as the data under Name and Date of Birth columns are identical to the ones in row 3, referring to the same customer.

Structured Data Data Preparation (Cleansing)

Exhibit 4. Data After Cleansing

1	ID	Name	Gender	Date of Birth	Salary	Other Income	State	Credit Card
2	1	Mr. ABC	M	12/5/1970	USD 50200	USD 5000	VA	Y
3	2	Ms. XYZ	F	1/15/1975	USD 60500	USD 0	NY	Y
4	3	Mr. EFG	M	1/13/1979	USD 65000	USD 1000	CA	N
5	6	Mr. GHI	M	9/10/1942	USD 0	USD 55000	TX	N
6	7	Mr. TUV	M	2/27/1956	USD 300000	USD 50000	CT	Y
7	8	Ms. DEF	F	4/4/1980	CAD 55000	CAD 0	British Columbia	N

Data Wrangling (Preprocessing)

To make structured data ready for analyses, the data should be preprocessed. Data preprocessing primarily includes transformations and scaling of the data. These processes are exercised on the cleansed dataset. The following transformations are common in practice:

- 1 Extraction: A new variable can be extracted from the current variable for ease of analyzing and using for training the ML model.** In Exhibit 4, the Date of Birth column consists of dates that are not directly suitable for analyses. Thus, an additional variable called “Age” can be extracted by calculating the number of years between the present day and date of birth.
- 2 Aggregation: Two or more variables can be aggregated into one variable to consolidate similar variables.** In Exhibit 4, the two forms of income, Salary and Other Income, can be summed into a single variable called Total Income.
- 3 Filtration: The data rows that are not needed for the project must be identified and filtered.** In Exhibit 4, row 7 (ID 8) has a non-US state; however, this dataset is for the US-based bank customers where it is required to have a US address.

Data Wrangling (Preprocessing)

4 Selection: The data columns that are intuitively not needed for the project can be removed.

This should not be confused with feature selection, which is explained later. In Exhibit 4, Name and Date of Birth columns are not required for training the ML model. The ID column is sufficient to identify the observations, and the new extracted variable Age replaces the Date of Birth column.

5 Conversion: The variables can be of different types: nominal, ordinal, continuous, and categorical. The variables in the dataset must be converted into appropriate types to further process and analyze them correctly.

This is critical for ML model training. Before converting, values must be stripped out with prefixes and suffixes, such as currency symbols. In Exhibit 4, Name is nominal, Salary and Income are continuous, Gender and Credit Card are categorical with 2 classes, and State is ordinal. In case row 7 is not excluded, the Salary in row 7 must be converted into US dollars. Also, the conversion task applies to adjusting time value of money, time zones, and others when present.

Data Wrangling (Preprocessing)

Outliers may be present in the data, and domain knowledge is needed to deal with them. Any outliers that are present must first be identified. The outliers then should be examined and a decision made to either remove or replace them with values imputed using statistical techniques.

Data Wrangling (Preprocessing)

In practice, several techniques can be used to **detect outliers in the data**.

Standard deviation can be used to identify outliers in normally distributed data. In general, a data value that is outside of 3 standard deviations from the mean may be considered an outlier.

The interquartile range (IQR) can be used to identify outliers in data with any form of distribution. IQR is the difference between the 75th and the 25th per - centile values of the data. The center of the IQR is the median (50th percentile). In general, data values outside of 1.5 IQR are considered as outliers and values outside of 3 IQR as extreme values.

Data Wrangling (Preprocessing)

There are several practical **methods for handling outliers**.

When extreme values and outliers are **simply removed** from the dataset, it is known as trimming (also called **truncation**). For example, a 5% trimmed dataset is one for which the 5% highest and the 5% lowest values have been removed.

When extreme values and outliers are **replaced with the maximum** (for large value outliers) and minimum (for small value outliers) values of data points that are not outliers, the process is known as **winsorization**.

Data Wrangling (Preprocessing)

Exhibit 5 Data After Applying Transformations

1	ID	Gender	Age	Total Income	State	Credit Card
2	1	M	48	55200	VA	Y
3	2	F	43	60500	NY	Y
4	3	M	39	66000	CA	N
5	6	M	76	55000	TX	N

Data Wrangling (Preprocessing)

Scaling is a process of **adjusting the range of a feature** by shifting and changing the scale of data. Variables, such as age and income, can have a diversity of ranges that result in a heterogeneous training dataset. For better ML model training when using such methods as support vector machines (SVMs) and artificial neural networks (ANNs), all variables should have values in the same range to make the dataset homogeneous.

It is important to remove outliers before scaling is performed. Here are two of the most common ways of scaling:

Data Wrangling (Preprocessing)

1 Normalization is the process of rescaling numeric variables in the range of [0, 1]. To normalize variable X, the minimum value (X_{\min}) is subtracted from each observation (X_i), and then this value is divided by the difference between the maximum and minimum values of X ($X_{\max} - X_{\min}$) as follows:

$$X_{i \text{ (normalized)}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

2 Standardization is the process of both centering and scaling the variables. Centering involves subtracting the mean (μ) of the variable from each observation (X_i) so the new mean is 0. Scaling adjusts the range of the data by dividing the centered values ($X_i - \mu$) by the standard deviation (σ) of feature X. The resultant standardized variable will have an arithmetic mean of 0 and standard deviation of 1.

$$X_{i \text{ (standardized)}} = \frac{X_i - \mu}{\sigma}$$

Data Wrangling (Preprocessing)

Normalization is sensitive to outliers, so treatment of outliers is necessary before normalization is performed. **Normalization can be used when the distribution of the data is not known.**

Standardization is relatively less sensitive to outliers as it depends on the mean and standard deviation of the data. However, the data must be normally distributed to use standardization.

Data exploration objectives and methods

Data exploration is a crucial part of big data projects. The prepared data are explored to investigate and comprehend data distributions and relationships. The knowledge that is gained about the data in this stage is used throughout the project. **The outcome and quality of exploration strongly affects ML model training results.**

Domain knowl - edge plays a vital role in exploratory analysis as this stage should involve cooperation between analysts, model designers, and experts in the particular data domain. **Data exploration without domain knowledge can result in ascertaining spurious relationships** among the variables in the data that can mislead the analyses

Data exploration involves **three important tasks**:

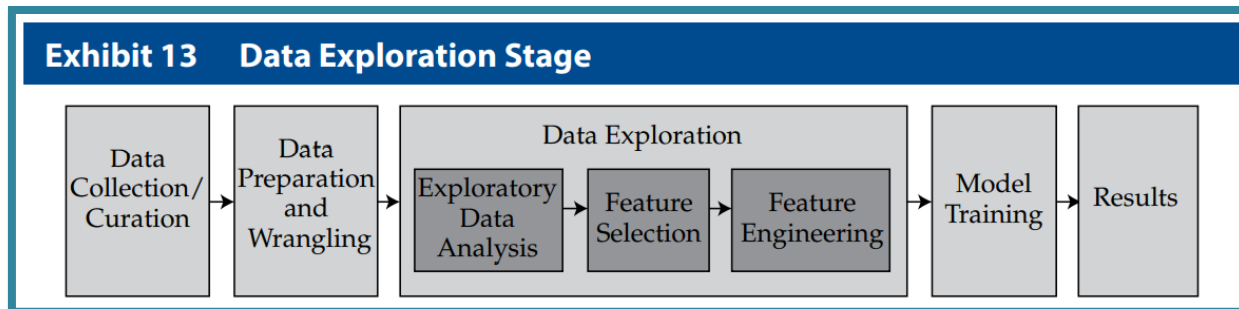
1. exploratory data analysis,
2. feature selection,
3. and feature engineering.

Data exploration objectives and methods

Exploratory data analysis (EDA) is the preliminary step in data exploration.

Exploratory **graphs, charts, and other visualizations, such as heat maps and word clouds**, are designed to **summarize and observe data**. In practice, many exploratory graphs are made for investigation and can be made swiftly using statistical programming and generic spreadsheet software tools. **Data can also be summarized and examined using quantitative methods, such as descriptive statistics and central tendency measures.**

An important objective of EDA is to serve as a communication medium among project stakeholders, including business users, domain experts, and analysts. Relatively quick and easy exploratory visualizations help stakeholders connect and ensure the prepared data are sensible.

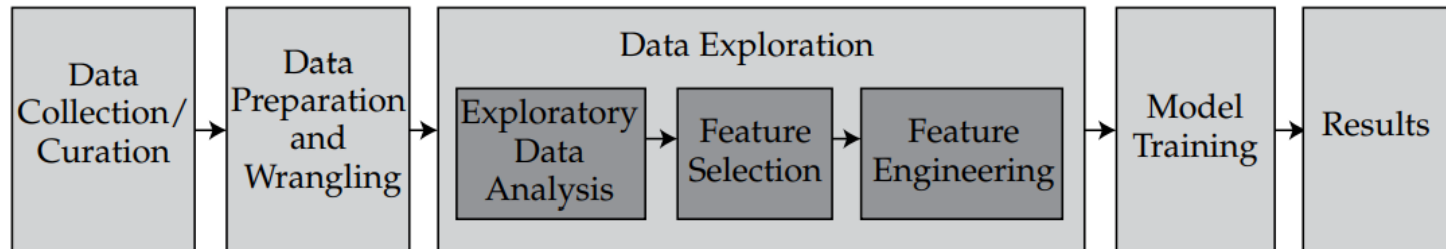


Data exploration objectives and methods

Other objectives of EDA include:

- understanding data properties,
- finding patterns and relationships in data,
- inspecting basic questions and hypotheses,
- documenting data distributions and other characteristics, and
- planning modeling strategies for the next steps

Exhibit 13 Data Exploration Stage

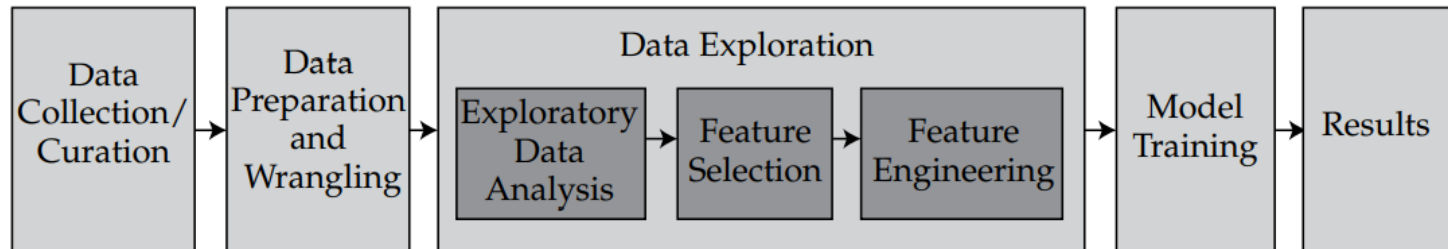


Data exploration objectives and methods

Feature selection is a process whereby only pertinent features from the dataset are selected for ML model training. Selecting fewer features decreases ML model complexity and training time.

Feature engineering is a process of creating new features by changing or transforming existing features. Model performance heavily depends on feature selection and engineering.

Exhibit 13 Data Exploration Stage



Exploratory data analysis for structured data

For structured data, each data table row contains an observation and each column contains a feature. **EDA can be performed on a single feature (one-dimension) or on multiple features (multi-dimension).** For high-dimension data with many features, EDA can be facilitated by using a dimension reduction technique, such as principal components analysis (PCA). Based on the number of dimensions, the exploratory techniques will vary.

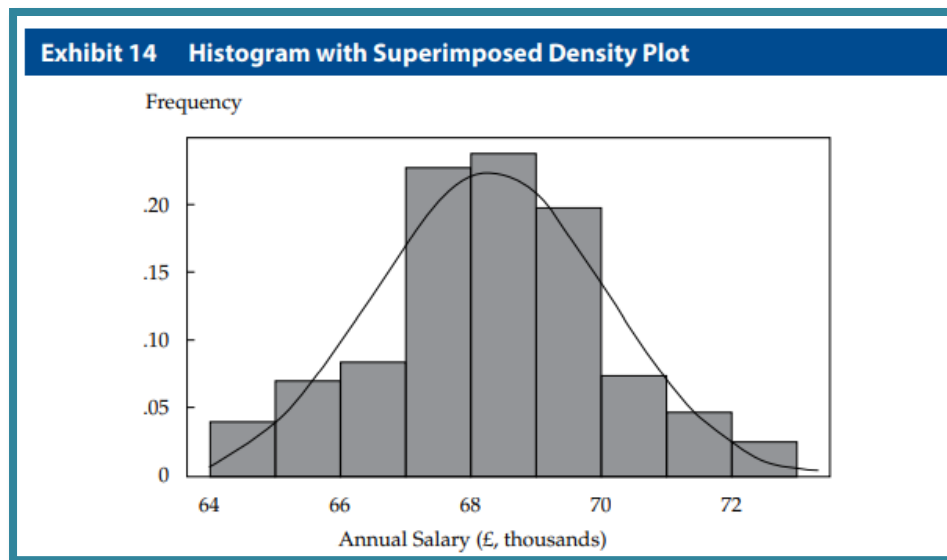
For one-dimensional data, summary statistics, such as mean, median, quartiles, ranges, standard deviations, skewness, and kurtosis, of a feature can be computed. One-dimension visualization summarizes each feature in the dataset. The basic onedimension exploratory visualizations are as follows: ■ Histograms ■ Bar charts ■ Box plots ■ Density plots

- Histograms represent equal bins of data and their respective frequencies. They can be used to understand the high-level distribution of the data.
- Bar charts summarize the frequencies of categorical variables.
- Box plots show the distribution of continuous data by highlighting the median, quartiles, and outliers of a feature that is normally distributed.

Exploratory data analysis for structured data

Density plots are another effective way to understand the distribution of continuous data.

Density plots are smoothed histograms and are commonly laid on top of histograms, as shown in Exhibit 14. This histogram shows a hypothetical annual salary distribution (in £) of entry-level analyst positions at UK banks. The data represent a normal distribution with an approximate mean of £68,500.



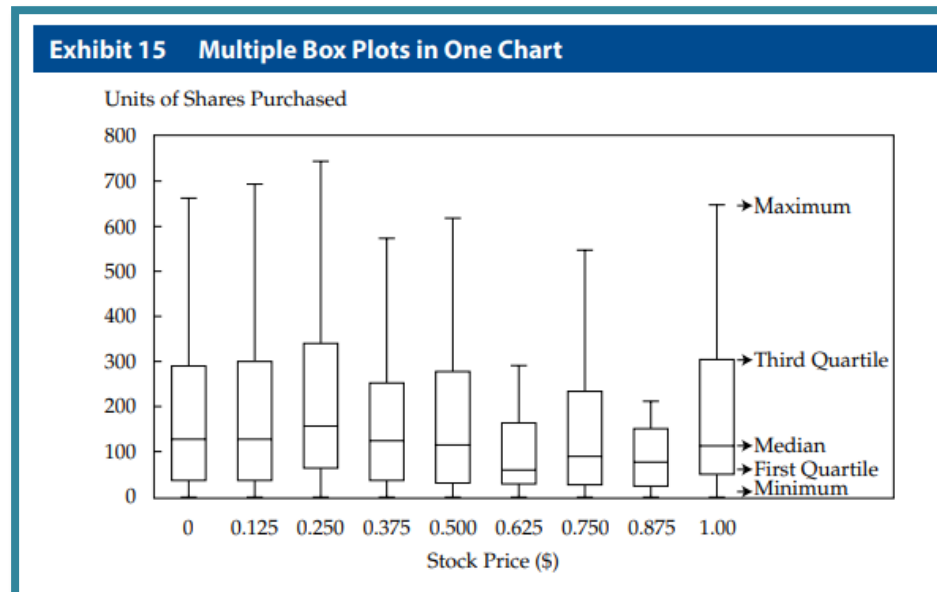
Exploratory Data Analysis for Structured Data

For data with two or more dimensions, summary statistics of relationships, such as a correlation matrix, can be calculated. Two- or more-dimensional visualization explores interactions between different features in the dataset. Common methods include **scatterplots and line graphs**. In multi-dimensional visualization, one-dimensional plots are overlaid to summarize each feature, thus enabling comparison between features. Additionally, attributes (e.g., color, shape, and size) and legends can be used creatively to pack more information about the data into fewer graphs.

For multivariate data, commonly utilized exploratory visualization designs include stacked bar and line charts, multiple box plots, and scatterplots showing multivariate data that use different colors or shapes for each feature. **Multiple box plots can be arranged in a single chart, where each individual box plot represents a feature. Such a multi-box plot chart assesses the relationship between each feature (x-axis) in the dataset and the target variable of interest (y-axis).**

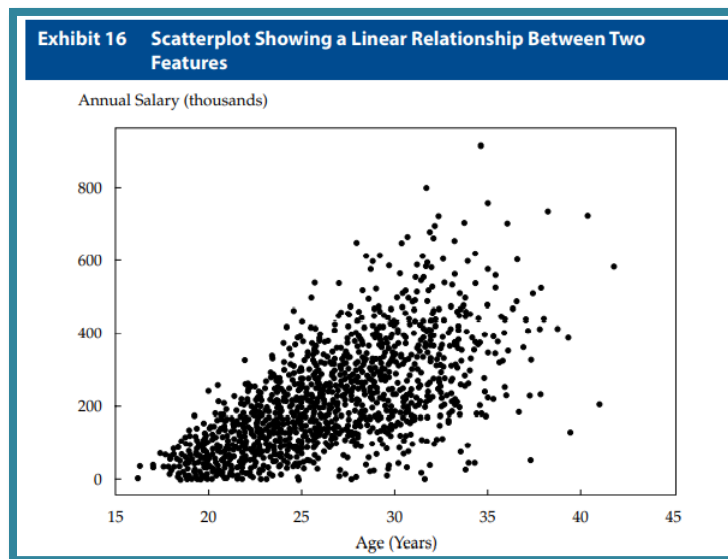
Exploratory Data Analysis for Structured Data

The multi-box plot chart in Exhibit 15 represents units of shares purchased versus stock price for a hypothetical stock. The x-axis shows the stock price in increments of \$0.125, and the y-axis shows units of shares purchased. The individual box plots indicate the distribution of shares purchased at the different stock prices. When the stock price is \$0.25, the median number of shares purchased is the highest; when the stock price is \$0.625, the median number of shares purchased is the lowest. However, visually it appears that the number of shares purchased at different stock prices is not significantly different.



Exploratory Data Analysis for Structured Data

Two-dimensional charts can summarize and approximately measure relationships between two or more features. An example scatterplot in Exhibit 16 shows the interaction of two hypothetical features: age (x-axis) and annual salary (y-axis). The feature on the y-axis tends to increase as the feature on the x-axis increases. This pattern appears true visually; however, it may not be a statistically significant relationship. A scatterplot provides a starting point where relationships can be examined visually. These potential relationships should be tested further using statistical tests. Common parametric statistical tests include ANOVA, t-test, and Pearson correlation. Common non-parametric statistical tests include chi-square and the Spearman rank-order correlation.



Exploratory Data Analysis for Structured Data

In addition to visualization, descriptive statistics are a good means to summarize data. Central tendency measures as well as minimum and maximum values for continuous data are useful. Counts and frequencies for categorical data are commonly employed to gain insight regarding the distribution of possible values. EDA is not only useful for revealing possible relationships among features or general trends in the data; it is also beneficial during the feature selection and engineering stages. These possible relationships and trends in the data may be used to suggest new features that, when incorporated into a model, may improve model training.

Features Selection

Structured data consist of features, represented by different columns of data in a table or matrix.

After using **EDA to discover relevant patterns in the data**, it is essential to **identify and remove unneeded, irrelevant, and redundant features**. Basic diagnostic testing should also be performed on features to identify redundancy, heteroscedasticity, and multi-collinearity.

The objective of the feature selection process is **to assist in identifying significant features** that when used in a model retain the important patterns and complexities of the larger dataset while requiring fewer data overall. This last point is important since computing power is not free (i.e., explicit costs and processing time).

Typically, structured data even after the data preparation step can contain features that do not contribute to the accuracy of an ML model or that negatively affect the quality of ML training. The most desirable outcome is a parsimonious model with fewer features that provides the maximum predictive power out-of-sample. Feature selection must not be confused with the data preprocessing steps during data preparation. Good feature selection requires an understanding of the data and statistics, and comprehensive EDA must be performed to assist with this step.

Bon courage !