Exploratory Data Analysis for Feature Selection in Machine Learning

Google Cloud

Contents

About this guide

The objective of this document is to provide comprehensive guidance on exploratory data analysis (EDA) from both an intuitive (that is, through visualization) and a rigorous (that is, statistical) analysis. This guide aims to consolidate the different stories of conducting proper EDA, data cleaning, and feature selection in ML projects in a comprehensive approach that can easily be reproduced, so as to serve as a standard reference. Practitioners from different backgrounds and with varying experience in ML will benefit from following the process outlined.

In detail, this guide provides practical information on:

- Deciding which analysis or explorations are expected to be performed, based on the datasets (and prediction target) at hand
- Performing the selected analysis, taking into consideration:
    - Rigorous data analysis, focusing on the relationship between features or between features and labels, with rigorous reasoning (theory)
        - Descriptive analysis of each attribute in a dataset for numerical, categorical, and textual attributes
        - Correlation analysis of two attributes (numerical versus numerical, numerical versus categorical, and categorical versus categorical) through qualitative and/or quantitative analysis
        - Time- and agent-based contextual analysis for a deeper understanding of the dataset
    - Visualizations that help provide an intuitive understanding of the analysis result
    - A survey of the existing tools that are most suitable
- Determining the appropriate feature processing, based on the analysis result and domain knowledge

A concrete use case is also presented for the Adult Census Income dataset that applies the analysis and visualizations introduced.

Note: Feature selection itself is a comprehensive topic that generally includes filtering (forward and backward) methods, wrapper methods, and embedded methods. The feature selection recommendations discussed in this guide belong to the family of filtering methods, and as such, they are the most direct and typical steps after EDA. We recommend that interested readers check the following review for a complete overview of feature selection.

1. Introduction

Machine learning (ML) projects typically start with a comprehensive exploration of the provided datasets. It is critical that ML practitioners gain a deep understanding of:

- **The properties of the data**: schema, statistical properties, and so on
- **The quality of the data**: missing values, inconsistent data types, and so on
- **The predictive power of the data**: for example, the correlation of features with the target

This process lays the groundwork for the subsequent feature selection and engineering steps, and it provides a solid foundation for building good ML models. It is often said that if ML is the rocket engine, then the fuel is the (high-quality) data we can feed to ML algorithms.

Exploratory data analysis (EDA), feature selection, and feature engineering are frequently considered together, and they are all important steps in the ML journey. How the results of proper EDA can influence the subsequent decisions is not a trivial question given the complexity of the data and the problems we are currently dealing with.

2. Statistical data analysis

This section outlines the different **statistical analyses** performed, the motivation behind them, and examples of each. The goal of these analyses is to determine the **quality of features** and their **predictive power** in contrast with target value or label. They provide a more comprehensive understanding of the data and should be the first step in studying any dataset, not just those for ML projects.

The exploration of the data is conducted from three different angles: **descriptive**, **correlative**, and **contextual**. Each type introduces information on the predictive power of the features and enables an informed decision based on the outcome of the analysis. The methodology and process outlined in this section lays the foundation for the decision process described in Section 4.

2.1 Descriptive analysis (univariate analysis)

Descriptive analysis (or univariate analysis) provides an understanding of the characteristics of each attribute of the dataset. It also offers important evidence for feature selection in a later state.

The following table lists the suggested analysis for attributes that are common, numerical, categorical, and textual.

Attribute type	Statistic/calculation	Details
Common	Data type	Attribute's data type
	Missing values	Percentage of missing values *Note: The statistics that follow in this table should, in general, exclude the detected missing values.*
Numerical	Quantile statistics	Q1, Q2, Q3, min, max, range, interquartile range
	Descriptive statistics	Mean, mode, standard deviation, median absolute deviation, kurtosis, skewness
	Distribution histogram	Based on the appropriate number of bins
Categorical	Cardinality	Number of unique values for the categorical attribute For example: in the case of gender, the number of unique values is generally two: male and female.
	Unique counts	Number of occurrences for each unique value of the categorical attribute
Textual	Tokens	Number of unique tokens
	DF/TF	Distribution of document frequency and term frequency with or without standard English stop words

Further analysis will provide a better understanding of the relationships between the dataset attributes. This is the aim of correlation analysis.

2.2 Correlation analysis (bivariate analysis)

Correlation analysis (or bivariate analysis) examines the relationship between two attributes, say X and *Y*, and determines whether the two are correlated. This analysis can be done from two perspectives for various possible combinations:

- **Qualitative analysis**. Computation of the descriptive statistics of dependent numerical or categorical attributes against each unique value of the independent categorical attribute. This perspective helps to intuitively understand the relationship between *X* and *Y*. Visualizations are often used together with qualitative analysis as a more intuitive way of presenting the result.

- **Quantitative analysis.** A quantitative test of the relationship between *X* and *Y*, based on a hypothesis-testing framework. This perspective provides a formal and mathematical methodology to quantitatively determine the existence and/or strength of relationship.

The motivation for performing correlation analysis is to help determine:

- Which attributes are not predictive, in terms of correlation with the target value. Special attention is usually needed for unpredictive attributes to reveal a stronger relationship with the target.
- Which attributes hold redundant information that can be replaced with derived attributes. Including them all might only serve to increase the resource demand and will not introduce any gain in the ML process.

2.2.1 Qualitative analysis

Qualitative analysis is a primarily exploratory analysis used to gain an understanding of underlying reasons, opinions, and motivations. It provides insights into the problem and helps to develop ideas or hypotheses for potential quantitative research.

The following table lists the statistical analyses that could be performed between two features of either categorical or numerical type. There is no qualitative analysis for numerical pairs listed here, which is usually done by a sampled scatter plot.

Attribute types	Analysis
Both categorical (*X, Y*)	Contingence table with unique counts of *X* (*Y*) per unique value of *Y* (*X*)
Categorical (*X*) versus numerical (*Y*)	Descriptive statistics or histogram of *Y* per unique value of *X*

Example

The following contingency table shows the relationship between sex and handedness.

Handedness / Sex	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Analysis: The proportion of right- and left-handedness between males and females is different. The data indicates that there might exist certain relationships between the attributes *sex* and *handedness* under the studied context.

Note: The correlation observed might come from the biasing of either the data collection or the underlying experiment design, the verification of which is not the focus of this guide.

As the example illustrates, qualitative analysis offers some insight into the relationship between the attributes. To confirm the statistical significance of the difference, quantitative analysis can be further applied.

2.2.2 Quantitative analysis

Quantitative analysis quantifies relationships by generating numerical data or data that can be transformed into usable statistics. In the case of correlation analysis, the quantitative analysis is done through a statistical hypothesis test.

A hypothesis is proposed for the statistical relationship between two attributes. This proposed hypothesis is then compared to an alternative, idealized null hypothesis (which proposes that there is no relationship between the two attributes).

The comparison is deemed statistically significant if the relationship between the attributes would be an unlikely realization of the null hypothesis, according to a threshold probability: the significance level.

X	Y	
	Categorical	Numerical
Categorical	Chi-square test Information gain	Student T-test ANOVA Logistic regression Discretize *Y* (left column)
Numerical	Student T-test ANOVA Logistic regression Discretize *X* (row above)	Correlation Linear Regression Discretize *Y* (left column) Discretize *X* (row above)

Example

The corresponding quantitative analysis that can be peformed on the attributes *sex* and *handedness* is the chi-square test. The following steps detail the process.

Hypothesis:

- H0: Sex and handedness are independent
- H1: Sex and handedness are not independent

First, compute the expected distribution, assuming that H0 (the null hypothesis) holds. Given the contingency table already provided, the "expected contingency table" under H0 can be computed as follows:

Handedness / Sex	Right-handed	Left-handed	Total
Male	100*0.52*0.87=45	100*0.52*0.13=7	52
Female	100*0.48*0.87=42	100*0.48*0.13=6	48
Total	87	13	100

Test the statistics.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(43-45)^2}{45} + \frac{(44-42)^2}{42} + \frac{(9-7)^2}{7} + \frac{(4-6)^2}{6} = 1.42$$

$$p_{value} = Pr_{DF=1}(\chi^2 > 1.42) = 0.233$$

If the p-value is less than a threshold (for example, 0.05), then the null hypothesis is rejected. However, the p-value here is larger than the typical threshold, which means that the null hypothesis cannot be rejected and, correspondingly, that there is no strong enough evidence showing that "sex and handedness are not independent." This actually gives a contradictory result to the observations made in the previous section. It also demonstrates the importance of quantitative analysis.

2.3 Contextual analysis

Because neither descriptive analysis nor correlation analysis requires context information, both are both generic enough to be performed on any (structured) dataset. To further understand or profile the given dataset and to gain more domain-specific insights, two generic contextual information-based analyses are recommended: **time based** and **agent based**.

It is expected that the quality of the dataset can be further verified based on the domain knowledge from the contextual analysis result.

2.3.1 Time-based analysis

In many real-world datasets, the timestamp (or similar time-related attributes) is one of the key pieces of contextual information. For example, operation logs for an online API service usually contain the time that the log was generated and/or the time when a logged event happened. Transaction logs for a retail company usually contain the time at which a transaction occurred. Observing and/or understanding the characteristics of the data along the time dimension, with various granularities, is essential to understanding the data generation process and ensuring data quality.

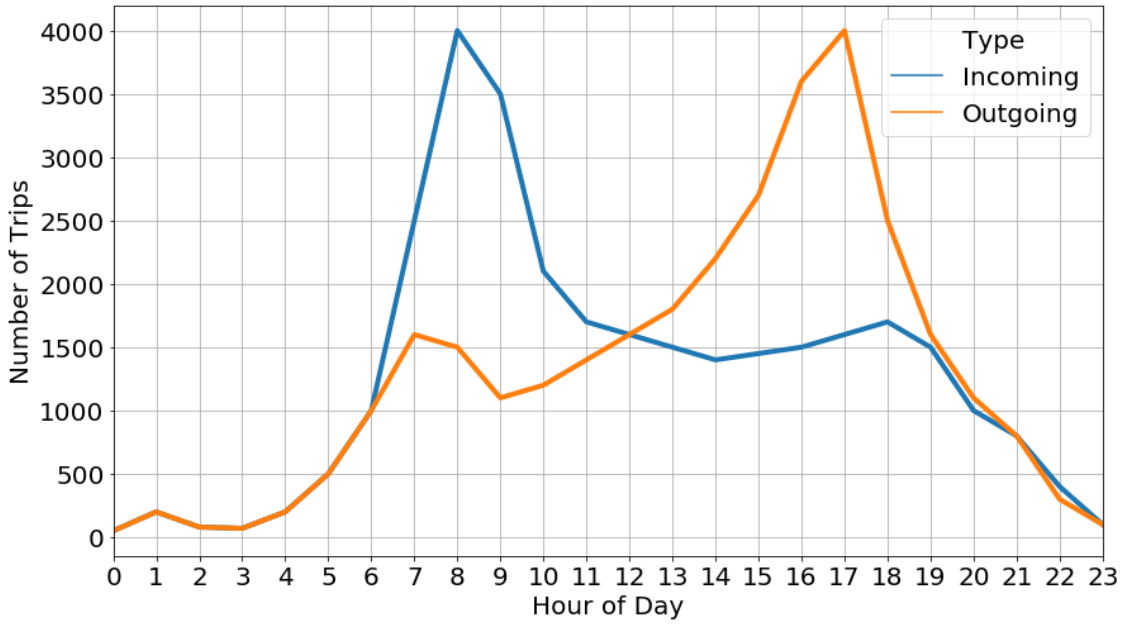
With the timestamp attribute, the following analyses could be performed:

- **Number of records** (transactions) per time interval
    - For example, hourly pageviews for a website
- **Number of unique values** (of an attribute) per time interval
    - For example, hourly unique visitors to a website
- **Descriptive statistics** per time interval
    - For example, the daily average session duration for a website

Note that the time interval could be in minutes, hours, or days, depending on the configuration.

Example

The following figure displays the average number of train trips per hour originating from and ending at one particular location based on a simulated dataset.



The conclusion that may be drawn is that peak times are around 8:30 a.m. and 5:30 p.m., which is consistent with the intuition that these are the times when people would typically leave home in the morning and return after a day of work.

2.3.2 Agent-based analysis

As an alternative to the timestamp, another common attribute is a unique identification (ID) for each record. These representing IDs provide important contextual information, including, for example:

- User ID in a transaction log of user activity
- Store ID and/or item ID in a transaction log for a retail shop
- Event ID in an operation log for a streaming system

WIth the ID attribute, the following analyses could be performed:

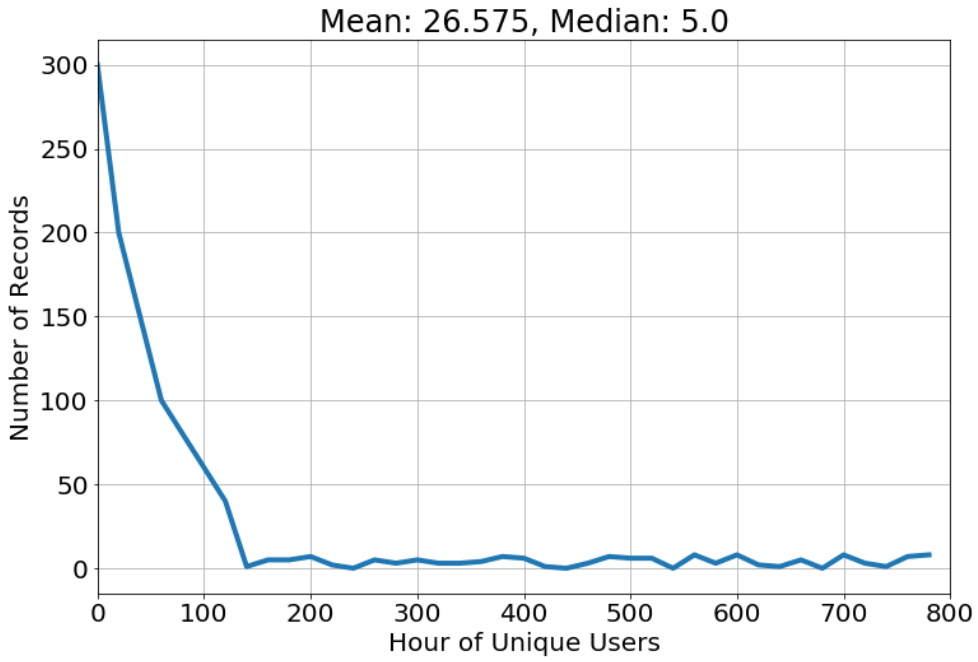
- Histogram and/or descriptive statistics of **number of records per agent**
    - For example, the average transaction per specific user in X months

- Histogram and/or descriptive statistics of **number of unique values** (of an attribute) per agent
    - For example, the average number of unique customers per specific store in X months

Example

This sample dataset was generated by simulating telco service–related activities (like call, SMS, and internet) for each mobile user in the format Timestamp, UserID, and EventID. The corresponding example statistics could be the number of calls made per user or the unique mobile numbers called per user.

The following figure is a histogram of the number of transactions (combined call, SMS, and internet) made per user in a week, where a user is uniquely identified by user ID.



In the figure, a long-tail distribution is observed with mean 26.6 and median 5. Domain knowledge can be used to check whether the distribution makes sense. If it does not, there could be issues at the data generating and/or collection stage.

3. Visualization for data analysis

Visualization presents data in a pictorial or graphical format. Such visualization is essential in gaining insight from the analysis result. The focus of this section is on listing and describing the various tools used to visualize the results of an analysis.

In the following table, possible matches are presented between analysis and visualizations. Treat this information as a recommendation guide, and do not restrict the innovation in how data can be visualized.

Visualization	Applicable analysis
Scatter plot	Linear and/or nonlinear correlation
Box plot	Quantile statistics [`Descriptive.Numerical`] Descriptive statistics of *Y* per unique value of *X* [`Correlation.Qualitative`]
Bar plot	Unique counts of categorical attribute [`Descriptive.Categorical`]
Line plot	Number of records (transactions) per time interval [`Contextual.Time`] Number of unique values (of an attribute) per time interval [`Contextual.Time`]
Histogram	Distribution of numerical attribute [`Descriptive.Numerical`] Distribution of number of records per agent [`Contextual.Agent`] Distribution of number of unique values (of an attribute) per agent [`Contextual.Agent`]
Heatmap	Unique counts of *X* per unique value of *Y* [`Correlation.Qualitative`]
Summary table	Descriptive statistics of numerical attribute [`Descriptive.Numerical`] Contingency table for unique counts of *X* per unique value of *Y* [`Correlation.Qualitative`] *Note: Summary tables are treated as a method of visualization.*

To learn how visualization can be used properly, see the comprehensive seaborn tutorial.

4. Feature selection and engineering

The ultimate goal of EDA (whether rigorous or through visualization) is to provide insights on the studied dataset. These insights can inspire the subsequent feature selection and engineering process. Listed in this section are the typical feature selections and engineering decisions that can be made, based on the analysis performed.

Note that the indicator of the issues discussed in the following subsections can be derived rigorously from statistical data analysis, and/or identified intuitively from visualization with domain knowledge judgment.

4.1 Feature selection based on descriptive analysis

Descriptive analysis provides the basic statistics of each attribute of the dataset, and based on this, some of the problematic features can be identified.

The following are indicators of problematic features, based on which feature selection decisions can be made accordingly.

- **High percentage of missing values.** The identified problem is that the attribute is missing in a significant proportion of the data points. The threshold can be set based on business domain knowledge.

  There are two options to handle this, depending on the business scenario:

  - The missing value, in certain contexts, is actually meaningful. For example, a missing value could indicate that a monitored, underlying process was not functioning properly. Therefore, **assigning a unique value** to the missing value records is a reasonable way to handle the situation.
  - A value can be missing due to misconfiguration, issues with data collection, or untraceable random reasons, in which case **the feature can simply be discarded** if the historic data cannot be reconstituted.

  More generally, missing values can be generally categorized into three cases:

  - **Missing at random**. The propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.
  - **Missing at completely random**. A certain missing value has nothing to do with its hypothetical value or with the values of other variables.

- **Missing not at random**. The missing value depends on the hypothetical value (for example, people with high salaries generally do not want to reveal their incomes in surveys) or the missing value is dependent on some other variable's value (for example, women generally don't want to reveal their age).

In the first two cases, it is safe to remove the data with missing values depending upon the percentage, while in the third case removing observations with missing values can produce a bias in the model.

- **Low variance of numeric attributes.** The identified problem is a very small variance of the feature compared to the typical value range of the feature, or the distribution is a sharp bell curve. In most cases, it is safe to remove numeric attributes with low variance. This will not harm the performance of the model, and it can reduce the complexity of the model.

- **Low entropy of categorical attributes.** The identified problem is a very small entropy of the feature, which means that most of the records have the same categorical values. In most cases, it is safe to remove categorical attributes with low entropy. This will not harm the performance of the model, and it can reduce the complexity of the model.

- **Imbalance of categorical target (class imbalance).** A dataset is said to be "highly class imbalanced" if a sample from one target class is significantly higher in number than others. This can be treated as a special case of the previous "low entropy of categorical attributes." In an imbalanced dataset, the class with a higher number of instances is called a *major class*, while those with relatively fewer instances are called *minor classes*.

  In this case, most of the classifiers are biased towards the major classes and, hence, display poor classification rates on minor classes. It is also possible that the classifier predicts everything as a major class and ignores the minor class. A variety of techniques have been proposed to handle class imbalance:

  - **Undersampling majority class.** The most important method in undersampling is the random undersampling method, which tries to balance the distribution of class by randomly removing the majority class sample. The main issue presented with this method is loss of valuable information.

  - **Oversampling minority class.** Random oversampling methods also help to achieve balance in class distribution by replicating the minority class sample. However, there is no new information added to the dataset. The synthetic minority oversampling technique (SMOTE) has been proposed for synthetic minority examples to oversample the minority class.

- **Alternative metric and/or loss function.** Accuracy is not the metric to use when working with an imbalanced dataset, and that can be misleading. There are metrics that have been designed to tell a more truthful story when working with imbalanced classes — for example, the area under the curve (AUC) and F1 score (average of precision and recall). It is also possible to alter the loss function by imposing an additional cost on the model for making classification mistakes on the minority class during training. These penalties can bias the model to pay more attention to the minority class.

- **Skew distribution.** The identified problem is that the distribution of the numeric attribute exhibits a long-tail shape.

  There are several options for handling this, depending on the business scenario:

  - Sometimes outliers can originate from incorrect data, in which case an understanding of the source of the error may allow changing the outlier value with a plausible one (for example, 99999 cigarettes per day, because the default value of the input zone is 99999).

  - If extreme values cannot be treated as outliers, transformations (such as log transformations) are usually recommended to generate a more balanced distribution.

  - Another transformation technique for handling the skew distribution is bucketization or binning. Bucketizing the numerical feature so that the number of data points in each bucket or bin is balanced can effectively mitigate the skew and preserve the predictive power. This is also called *equal-frequency binning*. There are other binning strategies (for example, equal-width binning) that cannot resolve the skew.

    An example of equal-frequency binning versus equal-width binning is:

    - Data: 0, 4, 12, 16, 16, 18, 24, 26, 28
    - Equal width
      i. Bin 1: 0, 4 [-, 10)
      ii. Bin 2: 12, 16, 16, 18 [10, 20)
      iii. Bin 3: 24, 26, 28 [20, +)
    - Equal frequency
      i. Bin 1: 0, 4, 12 [-, 14)
      ii. Bin 2: 16, 16, 18 [14, 21)
      iii. Bin 3: 24, 26, 28 [21, +)

- In certain cases, the extreme values are actually caused by outliers. Filtering out the extreme values or outliers would bring the distribution of the attribute back to normal. It is important to note that outlier removal should be applied consistently to both training and serving.

  There are two typical methods for removing the outliers:

  - Clipping the attribute at a computed percentile (for example, 99%), assuming that the majority of the data is valid and only a small percent (for example, 1%) is abnormal.
  - Filtering by fixed threshold, if it is known that there is a normal value range for the numerical attribute (for example, the latitude and longitude).

- **High cardinality.** The identified problem occurs when the number of unique values is too large for categorical attributes. This high cardinality creates problems for the typical one-hot-encoding process, creating a representation in an extremely high-dimensional space. Such high dimensionality often follows if the agent (user or item) ID is used as the feature.

  The typical remedy for this would be:

  - Applying a hash trick, which converts a high cardinality categorical attribute to a fixed sized one-hot-encoded space.
  - Embedding, a mapping from discrete objects, such as words, to vectors of real numbers.
  - Removing the feature if the number of occurrences per unique value of the attributes is too low (for example, less than 5), which can typically happen in cases like transaction ID or event ID.

4.2 Feature selection based on correlation analysis

The correlation analysis examines the relationship between two attributes. The typical action points triggered by the correlation analysis in the context of feature selection or feature engineering can be summarized as follows:

- **Low correlation between feature and target.** If the correlation between feature and target is found to be low, there are two possible reasons:
  - The feature is not useful in terms of predicting the desired target, and therefore it can be removed from the study.

- The feature is not useful given its available form, and transformation is required to reveal a stronger relationship with the target. One possible example is when the longitude and latitude don't provide clear information in the raw form. Bucketizing them to create the notion of "location" is often a useful transformation that could increase their correlation with the target.

For these reasons, before removing the features from the scope, it is recommended to have a domain expert review the analysis in order to make the final decision.

- **High correlation between features.** Another result that comes out of the correlation analysis and that requires special attention is the high correlation between features. Having highly correlated features could be a problem because:
  - No additional information is provided, nor does it help improve the model performance.
  - Computation requirements are increased during training.
  - Some models (linear, for example) may be made unstable.

  Typical methods for handling this issue include:
  - Removal of all but one of the highly correlated features.
  - Application of dimensionality reduction.
  - Use of a nonlinear model (for example, a neural network), which is usually more robust in the presence of correlated features.

4.3 Feature selection based on contextual analysis

The key purpose for providing contextual analysis is to help users gain a better understanding of the data. The visualization of the data across time and/or agent dimensions should help with understanding its context. However, the actions that can be taken in this case rely more heavily on the user's domain knowledge.

Thus, it is sufficient to simply present the visualization and leave the final decision to the user.

5. EDA tools ecosystem

A variety of tools exist to support the EDA phase of an ML project, though most of them don't support correlation analysis and further feature selection. For this reason, additional efforts are still needed. Some of the tools support code-free experiences and others are more evolved and require customized coding. Choose tools according to the specific requirements and your expertise.

This section provides a list of the existing publicly available tools and a comparison, with emphasis on Google Cloud Platform (GCP).

5.1 Existing tools

- Pandas profiling. This tool provides descriptive statistics and visual data exploration. It can also calculate correlations between numerical features.
- Facets. This data visualization tool for machine learning calculates data statistics and allows for data distribution and comparative feature visualization on both training and validation datasets.
- Cloud Dataprep. This tool is a cloud-based service built upon the Cloud Dataflow service that supports visually exploring, cleaning, and preparing data for analysis. It can generate features statistics and perform transformations.
- TensorFlow Data Validation. This tool provides calculation of summary statistics for the training and test datasets. It includes anomaly detection to identify anomalies (such as missing features, out-of-range values, or wrong feature types). It is focused on recurrent data validation in production pipelines rather than on initial data exploration.
- AutoML tables. This tool computes the basic statistics of each attribute of the imported dataset before the model is trained.
- Auto Data Exploration and Feature Recommendation Tool (Auto EDA). This tool automates the data analysis described in this guide, regardless of the scale of the data, using BigQuery as the backend compute engine. The result of the analysis is an automatically generated report presenting the findings in a compelling manner.

5.2 Feature comparison

The following table provides a comparison of the features of the various tools.

	Descriptive analysis	Correlation analysis	Contextual analysis	Information generated	Coding requirement
Auto EDA	No textual	Yes	No	Analysis report, in Markdown	Simple
Tensorflow Data Validation	No textual	No quantitative	No	Facet visualization	Advanced
Cloud Dataprep	No textual	No quantitative	No	Visualization on Cloud Dataprep UI	None
AutoML table	No textual	No	No	Visualization on AutoML Table UI	None
Facets	No textual	No quantitative	No	Facet visualization	Intermedia
Pandas profiling	Yes	Quantitative + Pearson correlation	No	Analysis report, in HTML	Simple

In summary, when selecting the appropriate exploration tool, consider the following:

- Existing development environment
- Team experience and expertise
- Size of the dataset
- Requirements of the analysis and visualization

6. Use case illustration

In this section, a concrete use case is presented, with an application of the analysis and visualizations introduced. The notebook is also available.

6.1 Dataset

The Adult Census Income dataset was extracted by Barry Becker from the 1994 census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). There are six numerical attributes and nine categorical attributes (including the target), with 32,561 rows.

The following tables display the categorical and numerical variables, respectively, in the sample dataset.

workclass	education	marital_status	occupation	relationship	race	sex	native_country	income_bracket
Private	9th	Married-civ-spouse	Other-service	Wife	Black	Female	United States	<=50K
Private	9th	Married-civ-spouse	Exec-managerial	Wife	Asian Pac Islander	Female	United States	>50K
Private	9th	Married-civ-spouse	Machine-op-inspct	Wife	White	Female	United States	>50K
Private	9th	Married-civ-spouse	Exec-managerial	Wife	White	Female	United States	<=50K
Private	9th	Married-civ-spouse	Tech-support	Wife	White	Female	United States	<=50K

age	functional_weight	education_num	capital_gain	capital_loss	hours_per_week
39	297847	5	3411	0	34
72	74141	5	0	0	48
45	178215	5	0	0	40
31	86958	5	0	0	40
55	176012	5	0	0	23

The objective for this use case is determining whether a person makes over 50K a year.

6.2 Descriptive analysis

Begin descriptive analysis on the Adult Census Income dataset by examining the data types of each attribute and locate any instances of missing values.

Numerical and categorical attributes are analyzed separately.

6.2.1 Data type and missing value

The following table provides information about the data types and missing values.

	Pandas_Dtype	python_type	Missing_Value	% Missing_Values
age	int64	int	0	0
workclass	object	str	0	0
functional_weight	int64	int	0	0
education	object	str	0	0
education_num	int64	int	0	0
marital_status	object	str	0	0
occupation	object	str	0	0
relationship	object	str	0	0
race	object	str	0	0
sex	object	str	0	0
capital_gain	int64	int	0	0
capital_loss	int64	int	0	0
hours_per_week	int64	int	0	0
native_country	object	str	0	0
income_bracket	object	str	0	0

There are no missing values for this dataset. The first column displays the Pandas data types. The second column provides the corresponding Python data types.

It's worth noting that a missing values check may not be as simple as looking for the value that's "missing." In some cases, missing values will be pre-filled by a certain fixed value (for example, "?" or "NA").

As such, it is recommended that you seek additional details about the data in question. Exploring categorical features is a viable option. For more information, see Section 6.2.3.

6.2.2 Numerical attributes

For numerical attributes, generate the following statistical information and histograms. There are different distributions of values for different numerical attributes from the histograms, and some of the problematic issues begin appearing.

For example, most of the people have 0 `capital_gain` and 0 `capital_loss`. This signals potential issues:

- A missing value may be recorded as 0.0
- Low variance and skew distribution
- Incorrect data

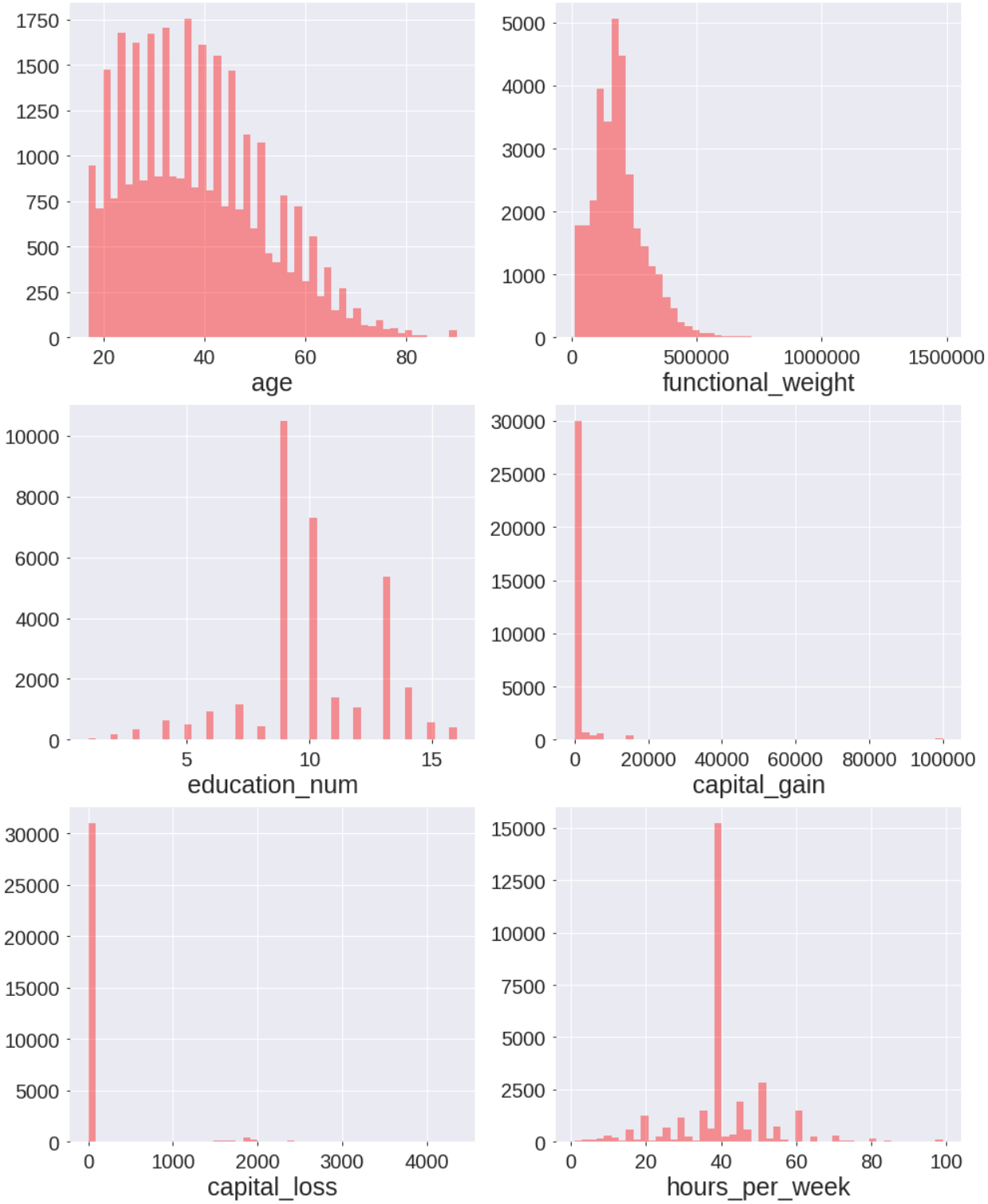
The next step is understanding the reason the pattern is generated and whether it can be used in the modeling with proper preprocessing.

The following table provides statistical information in descriptive analysis.

	max	range	IQR	mode	mad	kurtosis	skewness
age	90	73	20	36	11.189182	-0.166127	0.558743
functional_weight	1484705	1472420	119224	123011	77608.21854	6.218811	1.44698
education_num	16	15	3	9	1.903048	0.623444	-0.311676
capital_gain	99999	99999	0	0	1977.373437	154.799438	11.953848
capital_loss	4356	4356	0	0	166.462055	20.376802	4.594629
hours_per_week	99	98	5	40	7.583228	2.916687	0.227643

	mean	std	min	25%	50%	75%
age	38.581647	13.640433	17	28	37	48
functional_weight	189778.3665	105549.9777	12285	117827	178356	237051
education_num	10.080679	2.57272	1	9	10	12
capital_gain	1077.648844	7385.292085	0	0	0	0
capital_loss	87.30383	402.960219	0	0	0	0
hours_per_week	40.437456	12.347429	1	40	40	45

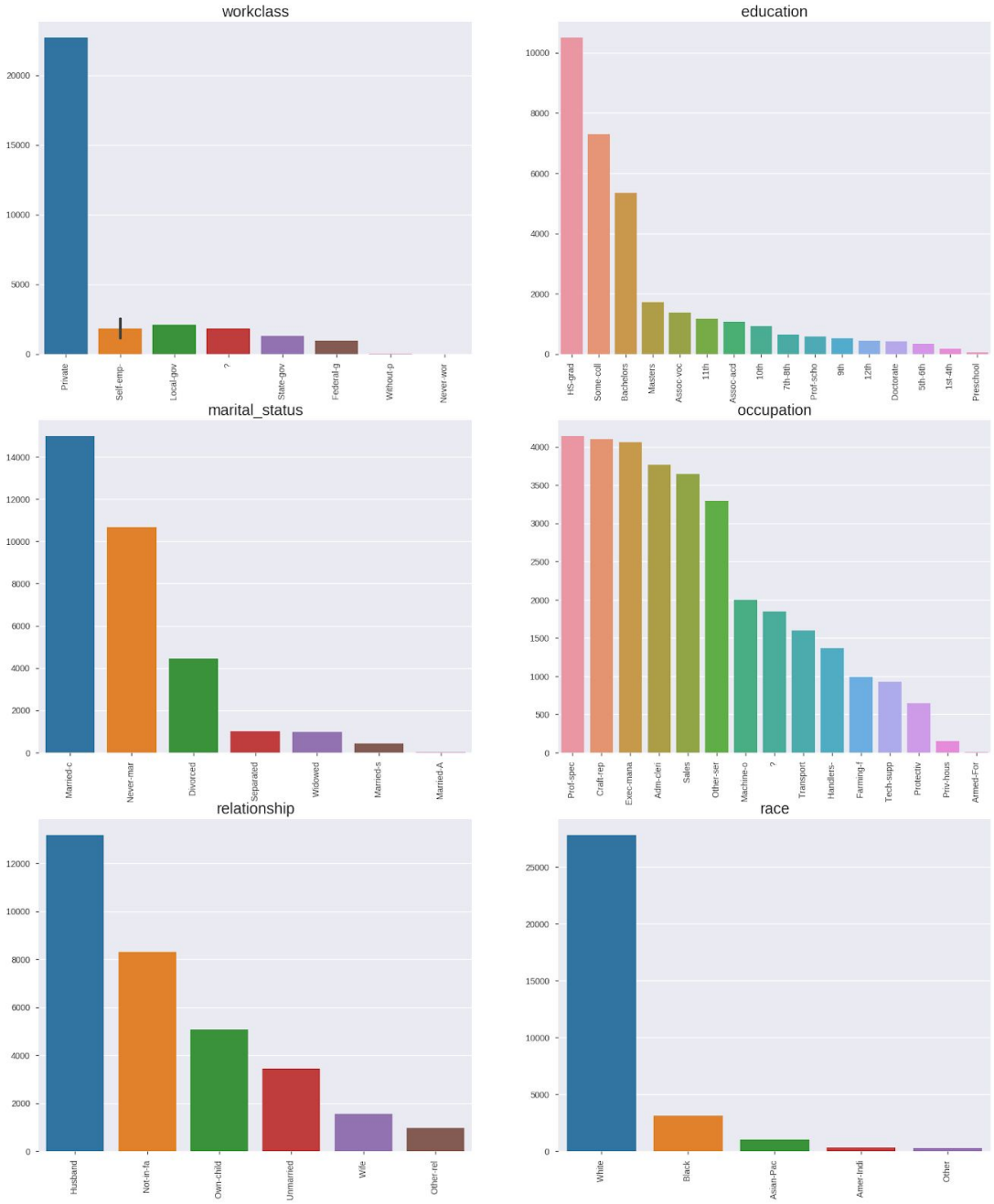
The following histograms display the numerical attributes.



Histograms of numerical attributes

6.2.3 Categorical attributes

For categorical attributes, the plots in the following figure can be generated. Note that for categories with lots of values, the labels may overlap. The focus is on the distribution of counts of each value (including uniform, bimodal, and long tail). An interesting discovery is the significant portion of the value "?" for the `workclass` feature. It is important to understand in what case "?" will be used. Moreover, we can observe that `race` has a potentially low entropy issue, which may need to be preprocessed before applying in the model. The following histograms present the selected categorical variables.



Histograms of categorical attributes

6.3 Correlation analysis

For qualitative and quantitative analysis, generate the tables in the following subsections with corresponding statistics. Note that this is not a full test of the data. Only a portion was chosen for the sake of demonstration.

6.3.1 Categorical versus categorical

Based on details displayed in the following contingency table, the proportions between male and female are significantly different for different races. The number of female working people and male working people are similar for the race `black`, while there are more than twice the amount of male working people as female working people for the race `white`. As such, a possible hypothesis could be that a large proportion of white females do not work after marriage.

	Female	Male	All
Amer-Indian-Eskimo	119	192	311
Asian-Pac-Islander	346	693	1039
Black	1555	1569	3124
Other	109	162	271
White	8642	19174	27816
All	10771	21790	32561

Using the following chi-square tests table as reference, observe that `income_bracket` is strongly correlated with the rest of the categorical attributes with a very small p-value. A possible hypothesis is that categorical attributes will have a positive impact on the prediction task.

cat1	cat2	chi_statistic	p_value	DoF
workclass	income_bracket	1045.7086	1.19E-210	18
education	income_bracket	4429.653302	0.00E+00	32
marital_status	income_bracket	6517.741654	0.00E+00	14
occupation	income_bracket	4031.97428	0.00E+00	30
relationship	income_bracket	6699.076897	0.00E+00	12
race	income_bracket	330.920431	4.43E-65	10
sex	income_bracket	1518.88682	0.00E+00	4
native_country	income_bracket	317.230386	8.56E-29	84

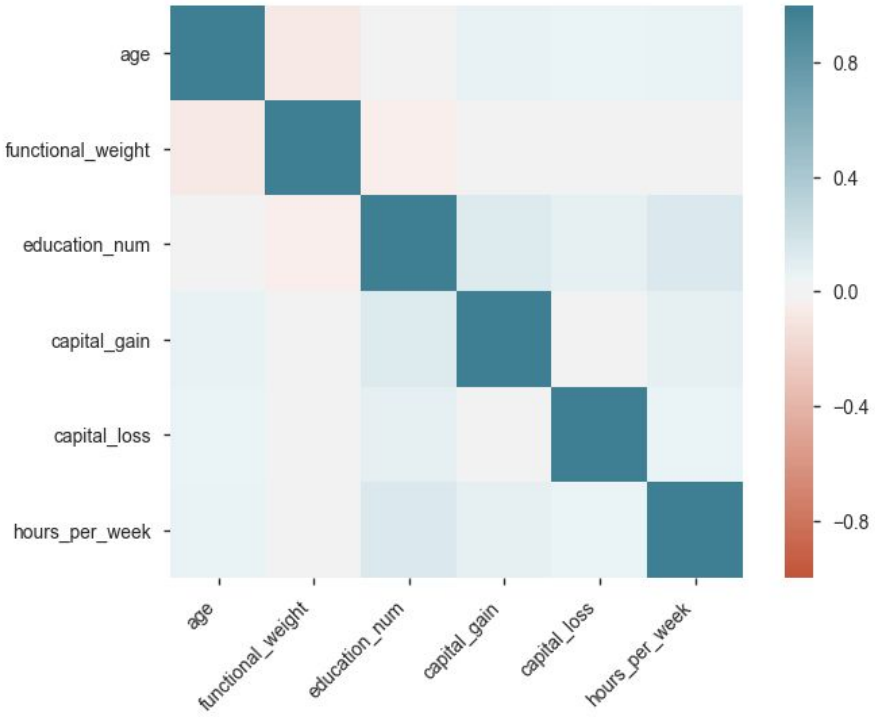
Use information gain to find the dependency of two categorical variables from a different perspective. Determine how much uncertainty for categorical variable A can be brought down by categorical variable B.

In the following table, which displays information gain, `income_bracket` becomes more certain if the person's `relationship` and `marital_status` are also known. Conversely, it does not seem to be affected if the `race` is known. Thus, it's expected that `relationship` and `marital_status` are significant features in predicting a person's income bracket.

cat-S	cat-A	H(S)	H(S|A)	IG
income_bracket	workclass	0.552011	0.537059	0.014952
income_bracket	education	0.552011	0.487139	0.064872
income_bracket	marital_status	0.552011	0.443514	0.108497
income_bracket	occupation	0.552011	0.487602	0.064409
income_bracket	relationship	0.552011	0.437388	0.114623
income_bracket	race	0.552011	0.546204	0.005807
income_bracket	sex	0.552011	0.526246	0.025765
income_bracket	native_country	0.552011	0.545984	0.006027

6.3.2 Numerical versus numerical

Using the following correlation heat map and table as reference, observe that there are no strong correlations between any two numerical variables, which signifies that these features may be able to provide complementary information when building the ML model to predict `income_bracket`.



Heatmap of Pearson correlation

	age	functional_ weight	education_ num	capital_ gain	capital_ loss	hours_ per_week
age	1	-0.076646	0.036527	0.077674	0.057775	0.068756
functional_weight	-0.076646	1	-0.043195	0.000432	-0.010252	-0.018768
education_num	0.036527	-0.043195	1	0.12263	0.079923	0.148123
capital_gain	0.077674	0.000432	0.12263	1	-0.031615	0.078409
capital_loss	0.057775	-0.010252	0.079923	-0.031615	1	0.054256
hours_per_week	0.068756	-0.018768	0.148123	0.078409	0.054256	1

6.3.3 Categorical versus numerical

The following table provides the T-test results between the target variable and numerical variables. Note that T-tests between the target variable and various numerical attributes show that `income_bracket` is strongly correlated with all numerical variables except for `functional_weight`, with a significance level set at 0.05.

Categorical	Value1	Value2	Numerical	p-value	t-statistic
income_bracket	<=50K	>50K	age	0.00E+00	-43.436244
income_bracket	<=50K	>50K	functional_weight	8.77E-02	1.707511
income_bracket	<=50K	>50K	education_num	0.00E+00	-64.187972
income_bracket	<=50K	>50K	capital_gain	0.00E+00	-41.341868
income_bracket	<=50K	>50K	capital_loss	2.69E-164	-27.474178
income_bracket	<=50K	>50K	hours_per_week	0.00E+00	-42.583873

Similarly, the next table, which displays the results of the ANOVA test between the target variable and numerical variables, also shows significant dependence between `income_bracket` and various categorical variables except for `functional_weight`. Results are consistent between the T-tests and ANOVA tests.

Categorical	Numerical	f-statistic	p-value
income_bracket	age	1886.707314	0.00E+00
income_bracket	functional_weight	2.915594	8.77E-02
income_bracket	education_num	4120.09578	0.00E+00
income_bracket	capital_gain	1709.150064	0.00E+00
income_bracket	capital_loss	754.830452	2.69E-164
income_bracket	hours_per_week	1813.386282	0.00E+00

In summary, most of the features observed are correlated with the target, which indicates the dataset should have significant predictive power.

Appendix

A. Hypothesis testing

A statistical hypothesis is an assumption about a population parameter that may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

The best way to determine whether a statistical hypothesis is true is by examining the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If the sample data is not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses:

- **Null hypothesis.** The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.
- **Alternative hypothesis.** The alternative hypothesis, denoted by H1 or Ha, is the hypothesis by which sample observations are influenced.

Statisticians follow a formal process to determine whether to reject the null hypothesis based on sample data. The following activities comprise its process, called *hypothesis testing*:

1. **State the hypotheses.** This involves stating the null and alternative hypotheses in such a way that they are mutually exclusive: that is, if one is true, the other must be false.

2. **Formulate an analysis plan.** The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.

3. **Analyze sample data.** Find the value of the test statistic (for example, its mean score, proportion, t statistic, or z-score) described in the analysis plan.

4. **Interpret results.** Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely based on the null hypothesis, reject the null hypothesis.

B. Pearson correlation coefficient

Mathematically, the sample Pearson correlation coefficient can be defined as:

$$r = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2 \sum_{i=1}^{N}(y_i-\bar{y})^2}}$$

where $r \in [-1, 1]$.

The Pearson correlation can be used as a ranking measurement for the **linear** fit of individual continuous variables:

- Correlations equal to 1 or –1 correspond to data points lying exactly on a line
- The coefficient is invariant under separate changes in location and scale in the two variables

Statistical inference:

$t = r\sqrt{\frac{n-2}{1-r^2}}$, with $df = n - 2$, where *r* is the sample Pearson correlation coefficient

C. Student T-test

The T-test is any statistical hypothesis test in which the test statistic follows a student's t-distribution under the null hypothesis.

A T-test can be applied for verifying the relationship between features or between feature and target variables in the following cases:

- Testing whether the distribution of the input variable between two groups (split against categorical variable) is the same
- Testing the null hypothesis that the true correlation coefficient ρ between two variables is equal to 0, based on the value of the sample correlation coefficient *r*

The following are examples of independent two-sample T-tests:

- H0: the difference between the two sample means is zero
- H1: the difference between the two sample means is not zero

Test statistics:

$t = \frac{\overline{x_A} - \overline{x_B}}{SE}$, where $SE = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$ and $df = min(n_A - 1,\ n_B - 1)$

D. Pearson's chi-square test

A statistical test can be applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. Pearson's chi-square test can be used to assess three types of comparisons:

- A **test of goodness of fit** establishes whether an observed frequency distribution differs from a theoretical distribution.
- A **test of homogeneity** compares the distribution of counts for two or more groups using the same categorical variable (for example, choice of activity — college, military, employment, travel — of graduates of a high school reported a year after graduation, sorted by graduation year, to see whether the number of graduates choosing a given activity has changed from class to class, or from decade to decade).
- A **test of independence** assesses whether unpaired observations on two variables, expressed in a contingency table, are independent of each other (for example, polling responses from people of different nationalities to see whether one's nationality is related to the response).

In the context of testing feature correlation, we apply the chi-square test to test the independence of two variables.

▶ *Statistical test*

Hypothesis:

- H0: feature and target are independent
- H1: feature and target are not independent

Test statistics:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- *r*: number of categories in the feature
- *c*: number of categories in target
- O_{ij} : number of instances with value *i* that are in class *j*
- E_{ij} : expected number of instances with value *i* and class *j*, where $E_{ij} = Np_ip_j$, $p_i = \frac{O_i}{N}$ and $p_j = \frac{O_j}{N}$, where N is the sample size

Under H0, $\chi^2 \sim \chi^2(d),\ df = (r-1)(c-1)$

* If the p-value is less than 0.05, then the null hypothesis is rejected; otherwise, the null hypothesis is accepted.

E. ANOVA statistical test

In the typical application of ANOVA, the null hypothesis is that all groups are random samples from the same population. For example, when studying the effect of different treatments on similar samples of patients, the null hypothesis would be that all treatments have the same effect (perhaps none). Rejecting the null hypothesis is taken to mean that the differences in observed effects between treatment groups are unlikely to be due to random chance. ANOVA is also commonly used to test the effectiveness of linear regression model statistical test.

Hypothesis:

- H0: $\mu_0 = \mu_1 = \mu_2 = ... = \mu_k$
- H1: $\mu_i \neq \mu_j$, where μ_i and μ_j are sample means of any two samples considered for the test

Test statistics:

$$F = \frac{MSG}{MSE}$$

Under H0, the *F* statistic has an *F* distribution with (*k*-1) and (*n*-*k*) degree of freedom in the numerator and denominator, where *k* is the number of groups and *n* is the number of data points within each group.

* If p-value is less than 0.05, then the null hypothesis is rejected; otherwise, the null hypothesis is accepted.

Source	SS	df	MS	F
Model/group	SSBG	*k*-1	$MSG = \frac{SSB}{k-1}$	$\frac{MSG}{MSE}$
Residual/error	SSE	*n*-*k*	$MSE = \frac{SSW}{n-k}$	
Total	SST	*n*-1		

SST= $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2$, sum of square total

SSBG= $\sum_{i=1}^{k} n_i(\overline{x}_i-\overline{x})^2$, between group variability between groups, where n_i is the number of data points in group i

SSWE= $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2$, within group variability within groups

SST=SSG+SSE

F. Information gain

A measure of the mutual dependence between the two variables, which can be computed as $Information\ Gain(A,\ S)\ = H(S) - H(S|A)$

- $H(S)$: entropy of attribute *S*
- $H(S|A)$: entropy of attribute *S* given the information of attribute *A*

The larger information gain indicates larger correlation between *A* and *S*.