

# Summer 2025 Data C100/C200 Final Reference Sheet

## Pandas

Suppose `df` is a DataFrame; `s` is a Series. `import pandas as pd`

Function	Description
<code>df.shape</code>	Returns a tuple containing the number of rows and columns, in that order
<code>df.index</code>	Returns the index (row labels) of <code>df</code> as an Index object
<code>df[col]</code>	Returns the column labeled <code>col</code> from <code>df</code> as a Series
<code>df.index[i]</code>	Returns the row label at position <code>i</code> from <code>df</code> 's index
<code>df[[col1, col2]]</code>	Returns a DataFrame containing the columns labeled <code>col1</code> and <code>col2</code>
<code>s.idxmax()</code>	Returns the index label of the first occurrence of the maximum value in Series <code>s</code>
<code>s.astype(dtype)</code>	Returns a Series casted to the specified type <code>dtype</code>
<code>s.loc[rows] / df.loc[rows, cols]</code>	Returns a Series/DataFrame with rows (and columns) selected by their index values
<code>s.iloc[rows] / df.iloc[rows, cols]</code>	Returns a Series/DataFrame with rows (and columns) selected by their positions
<code>s.isnull() / df.isnull()</code>	Returns boolean Series/DataFrame identifying missing values
<code>s.fillna(value) / df.fillna(value)</code>	Returns a Series/DataFrame where missing values are replaced by <code>value</code>
<code>s.isin(values) / df.isin(values)</code>	Returns a Series/DataFrame of booleans indicating if each element is in <code>values</code> .
<code>df.drop(labels, axis)</code>	Returns a DataFrame without the rows or columns named <code>labels</code> along <code>axis</code> (either 0 or 1)
<code>df.rename(index=None, columns=None)</code>	Returns a DataFrame with renamed columns from a dictionary <code>index</code> and/or <code>columns</code>
<code>df.sort_values(by, ascending=True)</code>	Returns a DataFrame where rows are sorted by the values in columns <code>by</code>
<code>s.sort_values(ascending=True)</code>	Returns a sorted Series
<code>s.unique()</code>	Returns a NumPy array of the unique values of <code>s</code> in the order that they appear
<code>s.value_counts()</code>	Returns the number of times each unique value appears in a Series
<code>pd.merge(left, right, how='inner', left_on=col1, right_on=col2)</code>	Returns a DataFrame joining <code>left</code> and <code>right</code> on columns labeled <code>col1</code> and <code>col2</code> ; the join is of type <code>inner</code>
<code>left.merge(right, left_on=col1, right_on=col2)</code>	Returns a DataFrame joining <code>left</code> and <code>right</code> on columns labeled <code>col1</code> and <code>col2</code>
<code>df.pivot_table(values=None, index=None, columns=None, aggfunc='mean', fill_value=None)</code>	Returns a DataFrame pivot table where columns are unique values from <code>columns</code> (column name or list), and rows are unique values from <code>index</code> (column name or list); cells are collected <code>values</code> using <code>aggfunc</code> . If <code>values</code> is not provided, cells are collected for each remaining column with multi-level column indexing.
<code>df.set_index(col)</code>	Returns a DataFrame that uses the values in the column labeled <code>col</code> as the row index
<code>df.reset_index()</code>	Returns a DataFrame that has row index 0, 1, etc., and adds the current index as a column

Let `grouped = df.groupby(by)` where `by` can be a column label or a list of labels

Function	Description
<code>grouped.count()</code>	Return a DataFrame containing the size of each group, excluding missing values
<code>grouped.size()</code>	Return a Series containing size of each group, including missing values
<code>grouped.mean()/min()/max()</code>	Return a Series/DataFrame containing mean/min/max of each group for each column, excluding missing values
<code>grouped.first()/last()</code>	Return a Series/DataFrame containing first/last entry of each group for each column, excluding missing values
<code>grouped.filter(f)</code> <code>grouped.agg(f)</code>	Filters or aggregates using the given function <code>f</code>
Function	Description
<code>s.str.len()</code>	Returns a Series containing length of each string

Function	Description
<code>s.str[a:b]</code>	Returns a Series where each element is a slice of the corresponding string indexed from <code>a</code> (inclusive, optional) to <code>b</code> (non-inclusive, optional)
<code>s.str.lower()/s.str.upper()</code>	Returns a Series of lowercase/uppercase versions of each string
<code>s.str.replace(pat, repl, regex=False)</code>	Returns a Series that replaces occurrences of substrings matching <code>pat</code> with string <code>repl</code> . When <code>regex=False</code> , <code>pat</code> is treated as a literal string; when <code>regex=True</code> , <code>pat</code> is treated as a RegEx pattern.
<code>s.str.contains(pat)</code>	Returns a boolean Series indicating if a substring matching the regex <code>pat</code> is contained in each string
<code>s.str.extract(pat)</code>	Returns a DataFrame of the first subsequence of each string that matches the regex <code>pat</code> . If <code>pat</code> contains one group, then only the substring matching the group is extracted
<code>s.str.split(pat=" ")</code>	Splits the strings in <code>s</code> at the delimiter <code>pat</code> (defaults to a whitespace). Returns a Series of lists, where each list contains strings of the characters before and after the split.

## Visualization

Matplotlib: `x` and `y` are sequences of values. `import matplotlib.pyplot as plt`

Function	Description
<code>plt.plot(x, y)</code>	Creates a line plot of <code>x</code> against <code>y</code>
<code>plt.scatter(x, y)</code>	Creates a scatter plot of <code>x</code> against <code>y</code>
<code>plt.hist(x, bins=None)</code>	Creates a histogram of <code>x</code> ; <code>bins</code> can be an integer or a sequence
<code>plt.bar(x, height)</code>	Creates a bar plot of categories <code>x</code> and corresponding heights <code>height</code>

Seaborn: `x` and `y` are column names in a DataFrame `data`. `import seaborn as sns`

Function	Description
<code>sns.countplot(data=None, x=None)</code>	Create a barplot of value counts of variable <code>x</code> from <code>data</code>
<code>sns.histplot(data=None, x=None, stat='count', kde=False)</code> <code>sns.displot(data=None, x=None, kind='hist', rug=False)</code>	Creates a histogram of <code>x</code> from <code>data</code> , where bin statistics <code>stat</code> is one of <code>'count'</code> , <code>'frequency'</code> , <code>'probability'</code> , <code>'percent'</code> , and <code>'density'</code> ; optionally overlay a kernel density estimator. <code>displot</code> is similar but can optionally overlay a rug plot and/or a KDE plot
<code>sns.rugplot(data=None, x=None)</code>	Adds a rug plot on the x-axis of variable <code>x</code> from <code>data</code>
<code>sns.boxplot(data=None, x=None, y=None)</code> <code>sns.violinplot(data=None, x=None, y=None)</code>	Create a boxplot of a numeric feature (e.g., <code>y</code> ), optionally factoring by a category (e.g., <code>x</code> ), from <code>data</code> . <code>violinplot</code> is similar but also draws a kernel density estimator of the numeric feature
<code>sns.scatterplot(data=None, x=None, y=None)</code>	Create a scatterplot of <code>x</code> versus <code>y</code> from <code>data</code>
<code>sns.lmplot(data=None, x=None, y=None, fit_reg=True)</code>	Create a scatterplot of <code>x</code> versus <code>y</code> from <code>data</code> , and by default overlay a least-squares regression line
<code>sns.jointplot(data=None, x=None, y=None, kind='scatter')</code>	Combine a bivariate scatterplot of <code>x</code> versus <code>y</code> from <code>data</code> , with univariate density plots of each variable overlaid on the axes; <code>kind</code> determines the visualization type for the distribution plot, can be <code>scatter</code> , <code>kde</code> or <code>hist</code>

## Regular Expressions

Operator	Description	Operator	Description
<code>.</code>	Matches any character except <code>\n</code>	<code>*</code>	Matches preceding character/group zero or more times
<code>\</code>	Escapes metacharacters	<code>?</code>	Matches preceding character/group zero or one times
<code> </code>	Matches expression on either side of expression; has lowest priority of any operator	<code>+</code>	Matches preceding character/group one or more times
<code>\d, \w, \s</code>	Predefined character group of digits (0-9), alphanumerics (a-z, A-Z, 0-9, and underscore), or whitespace, respectively	<code>^, \$</code>	Matches the beginning and end of the line, respectively
<code>\D, \W, \S</code>	Inverse sets of <code>\d, \w, \s</code> , respectively	<code>( )</code>	Capturing group used to create a sub-expression

Operator	Description	Operator	Description
<code>{m}</code>	Matches preceding character/group exactly <code>m</code> times	<code>[ ]</code>	Character class used to match any of the specified characters or range (e.g. <code>[abcde]</code> is equivalent to <code>[a-e]</code> )
<code>{m, n}</code>	Matches preceding character/group at least <code>m</code> times and at most <code>n</code> times. If either <code>m</code> or <code>n</code> are omitted, set lower/upper bounds to 0 and $\infty$ , respectively	<code>[^ ]</code>	Invert character class; e.g. <code>[^a-c]</code> matches all characters except <code>a</code> , <code>b</code> , <code>c</code>

Modified lecture example for capture groups:

```
import re
lines = '169.237.46.168 - - [26/Jan/2014:10:47:58 -0800] "GET ... HTTP/1.1"'
re.findall(r'\[(\d+\/(\w+)\.\/\d+:\d+:\d+ .+)\]', lines) # returns ['Jan']
```

Function	Description
<code>re.match(pattern, string)</code>	Returns a match if zero or more characters at beginning of <code>string</code> matches <code>pattern</code> , else None
<code>re.search(pattern, string)</code>	Returns a match if zero or more characters anywhere in <code>string</code> matches <code>pattern</code> , else None
<code>re.findall(pattern, string)</code>	Returns a list of all non-overlapping matches of <code>pattern</code> in <code>string</code> (if none, returns empty list)
<code>re.sub(pattern, repl, string)</code>	Returns <code>string</code> after replacing all occurrences of <code>pattern</code> with <code>repl</code>

Modeling

Concept	Formula	Concept	Formula
Variance, $\sigma_x^2$	$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$	Correlation $r$	$r = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{\sigma_x} \frac{y_i - \bar{y}}{\sigma_y}$
$L_1$ loss	$L_1(y, \hat{y}) =  y - \hat{y} $	Linear regression estimate of $y$	$\hat{y} = \theta_0 + \theta_1 x$
$L_2$ loss	$L_2(y, \hat{y}) = (y - \hat{y})^2$	Least squares linear regression	$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \qquad \hat{\theta}_1 = r \frac{\sigma_y}{\sigma_x}$

Empirical risk with loss  $L$

$$R(\theta) = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{y}_i)$$

Multiple Linear Regression Formulas

Concept	Formula	Concept	Formula
Mean squared error	$R(\theta) = \frac{1}{n}   \mathbb{Y} - \mathbb{X}\theta  _2^2$	Normal equation	$\mathbb{X}^T \mathbb{X} \hat{\theta} = \mathbb{X}^T \mathbb{Y}$
Least squares estimate, if $\mathbb{X}$ is full rank	$\hat{\theta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$	Multiple $R^2$ (coefficient of determination)	$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y}$
Ridge Regression L2 Regularization	$\frac{1}{n}   \mathbb{Y} - \mathbb{X}\theta  _2^2 + \alpha   \theta  _2^2$	Squared L2 Norm of $\theta \in \mathbb{R}^d$	$  \theta  _2^2 = \sum_{j=1}^d \theta_j^2$
Ridge regression estimate (closed form)	$\hat{\theta}_{\text{ridge}} = (\mathbb{X}^T \mathbb{X} + n\alpha I)^{-1} \mathbb{X}^T \mathbb{Y}$	L1 Norm of $\theta \in \mathbb{R}^d$	$  \theta  _1 = \sum_{j=1}^d  \theta_j $
LASSO Regression L1 Regularization	$\frac{1}{n}   \mathbb{Y} - \mathbb{X}\theta  _2^2 + \alpha   \theta  _1$		

# Scikit-Learn

Package: <code>sklearn.linear_model</code>			
Linear Regression	Logistic Regression	Function(s)	Description
✓	-	<code>LinearRegression(fit_intercept=True)</code>	Returns an ordinary least squares Linear Regression model.
-	✓	<code>LogisticRegression(fit_intercept=True, penalty='l2', C=1.0)</code>	Returns an ordinary least squares Linear Regression model. Hyperparameter C is inverse of regularization parameter, C = 1/λ.
✓	-	<code>LassoCV()</code> , <code>RidgeCV()</code>	Returns a Lasso (L1 Regularization) or Ridge (L2 regularization) linear model, respectively, and picks the best model by cross validation.
✓	✓	<code>model.fit(X, y)</code>	Fits the scikit-learn <code>model</code> to the provided <code>x</code> and <code>y</code> .
✓	✓	<code>model.predict(X)</code>	Returns predictions for the <code>x</code> passed in according to the fitted <code>model</code> .
✓	✓	<code>model.predict_proba(X)</code>	Returns predicted probabilities for the <code>x</code> passed in according to the fitted <code>model</code> . If binary classes, will return probabilities for both class 0 and 1.
✓	✓	<code>model.coef_</code>	Estimated coefficients for the linear model, not including the intercept term.
✓	✓	<code>model.intercept_</code>	Bias/intercept term of the linear model. Set to 0.0 if <code>fit_intercept=False</code> .
Package: <code>sklearn.model_selection</code>			
		Function	Description
		<code>train_test_split(*arrays, test_size=0.2)</code>	Returns two random subsets of each array passed in, with 0.8 of the array in the first subset and 0.2 in the second subset.

## Probability

Let  $X$  have a discrete probability distribution  $P(X = x)$ .  $X$  has expectation  $\mathbb{E}[X] = \sum_x xP(X = x)$  over all possible values  $x$ , variance  $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ , and standard deviation  $\text{SD}(X) = \sqrt{\text{Var}(X)}$ .

Notes	Property of Expectation	Property of Variance
$X$ is a random variable.	$\mathbb{E}[X] = \sum_x xP(X = x)$	$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = E[X^2] - (E[X])^2$
$X$ is a random variable, $a, b \in \mathbb{R}$ are scalars.	$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$	$\text{Var}(aX + b) = a^2\text{Var}(X)$
$X, Y$ are random variables.	$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$	$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
$X$ is a Bernoulli random variable that takes on value 1 with probability $p$ and 0 otherwise.	$\mathbb{E}[X] = p$	$\text{Var}(X) = p(1 - p)$

## Parameter Estimation and Gradient Descent Update Rule

### Parameter Estimation

Suppose for each individual with fixed input  $x$ , we observe a random response  $Y = g(x) + \epsilon$ , where  $g$  is the true relationship and  $\epsilon$  is random noise with zero mean and variance  $\sigma^2$ .

For a new individual with fixed input  $x$ , define our random prediction  $\hat{Y}(x)$  based on a model fit to our observed sample  $(\mathbb{X}, \mathbb{Y})$ . The model risk is the mean squared prediction error between  $Y$  and  $\hat{Y}(x)$ :  $\mathbb{E}[(Y - \hat{Y}(x))^2] = \sigma^2 + \left(\mathbb{E}[\hat{Y}(x)] - g(x)\right)^2 + \text{Var}(\hat{Y}(x))$ .

Suppose that input  $x$  has  $p$  features and the true relationship  $g$  is linear with parameter  $\theta \in \mathbb{R}^{p+1}$ . Then  $Y = f_\theta(x) = \theta_0 + \sum_{j=1}^p \theta_j x_j + \epsilon$  and  $\hat{Y} = f_{\hat{\theta}}(x)$  for an estimate  $\hat{\theta}$  fit to the observed sample  $(\mathbb{X}, \mathbb{Y})$ .

### Gradient Descent Update Rule

For a learning rate  $\alpha$ , the gradient update step is:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla_{\theta} L(\theta^{(t)}, \mathbb{X}, \mathbb{Y})$$

where  $\nabla_{\theta} L(\theta^{(t)}, \mathbb{X}, \mathbb{Y})$  is the partial derivative/gradient of  $L$  with respect to  $\theta$ , evaluated at  $\theta^{(t)}$ .

## SQL

SQL syntax:

```
SELECT [DISTINCT]
    { * | expr [[AS] c_alias]
    {, expr [[AS] c_alias] ... } }
FROM tableref {, tableref}
[[INNER | LEFT ] JOIN table_name
    ON qualification_list]
[WHERE search_condition]
[GROUP BY colname {, colname...}]
[HAVING search_condition]
[ORDER BY column_list]
[LIMIT number]
[OFFSET number of rows];
```

Syntax	Description
SELECT column_expression_list	List is comma-separated. Column expressions may include aggregation functions ( <b>MAX</b> , <b>FIRST</b> , <b>COUNT</b> , <b>AVG</b> , etc). <b>AS</b> renames columns. <b>DISTINCT</b> selects only unique rows.
FROM s INNER JOIN t ON cond	Inner join tables <b>s</b> and <b>t</b> using <b>cond</b> to filter rows; the <b>INNER</b> keyword is optional.
FROM s LEFT JOIN t ON cond	Left outer join of tables <b>s</b> and <b>t</b> using <b>cond</b> to filter rows.
FROM s, t	Cross join of tables <b>s</b> and <b>t</b> : all pairs of a row from <b>s</b> and a row from <b>t</b>
WHERE a IN cons_list	Select rows for which the value in column <b>a</b> is among the values in a <b>cons_list</b> .
ORDER BY RANDOM LIMIT n	Draw a simple random sample of <b>n</b> rows.
ORDER BY a, b DESC	Order by column <b>a</b> (ascending by default) , then <b>b</b> (descending).
CASE WHEN pred THEN cons ELSE alt END	Evaluates to <b>cons</b> if <b>pred</b> is true and <b>alt</b> otherwise. Multiple <b>WHEN/THEN</b> pairs can be included, and <b>ELSE</b> is optional.
WHERE s.a LIKE 'p'	Matches each entry in the column <b>a</b> of table <b>s</b> to the text pattern <b>p</b> . The wildcard <b>%</b> matches at least zero characters.
LIMIT number	Keep only the first <b>number</b> rows in the return result.
OFFSET number	Skip the first <b>number</b> rows in the return result.

## Singular Value Decomposition (SVD)

The SVD of a matrix  $X$  is  $X = USV^T$ .

Syntax	Description
np.linalg.svd(X, full_matrices = True)	SVD of <b>x</b> with shape ( <b>M</b> , <b>N</b> ) that returns <b>u</b> , <b>s</b> , <b>vt</b> , where <b>s</b> is a 1D array of <b>x</b> 's singular values. If <b>full_matrices=True</b> , <b>u</b> and <b>vt</b> have shapes ( <b>M</b> , <b>M</b> ) and ( <b>N</b> , <b>N</b> ) respectively; otherwise shapes are ( <b>M</b> , <b>K</b> ) and ( <b>K</b> , <b>N</b> ), respectively, where <b>K</b> = min( <b>M</b> , <b>N</b> ).

## Classification and Logistic Regression

### Classification Performance

Suppose you predict  $n$  datapoints.

Metric	Formula	Other Names
Accuracy	$\frac{TP+TN}{n}$	
Precision	$\frac{TP}{TP+FP}$	

Metric	Formula	Other Names
Recall/TPR	$\frac{TP}{TP+FN}$	True Positive Rate, Sensitivity
FPR	$\frac{FP}{FP+TN}$	False Positive Rate, $FPR = 1 - \text{Specificity}$

**Logistic Regression Model:** For input feature vector  $x$ ,  $\hat{P}_\theta(Y = 1|x) = \sigma(x^T\theta)$ , where  $\sigma(z) = 1/(1 + e^{-z})$ . For a single datapoint, define cross-entropy loss as  $-[y\log(p) + (1 - y)\log(1 - p)]$ , where  $p$  is the probability that the response is 1.

## Clustering

A datapoint's **silhouette score**  $S$  is defined as  $S = (B - A)/\max(A, B)$ , where  $A$  is the mean distance to other points in its cluster, and  $B$  is the mean distance to points in its closest cluster.