

# STAT 4710J Midterm Reference Sheet

## Pandas

Suppose `df` is a DataFrame; `s` is a Series. `pd` is the Pandas package.

Function	Description
<code>df[col]</code>	Returns the column labeled <code>col</code> from <code>df</code> as a Series.
<code>df[[col1, col2]]</code>	Returns a DataFrame containing the columns labeled <code>col1</code> and <code>col2</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>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
<code>s.value_counts()</code>	Returns the number of times each unique value appears in a Series
<code>pd.merge(left, right, how='inner', on='a')</code>	Returns a DataFrame joining DataFrames <code>left</code> and <code>right</code> on the column labeled <code>a</code> ; the join is of type <code>inner</code>
<code>left.merge(right, left_on=col1, right_on=col2)</code>	Returns a DataFrame joining DataFrames <code>left</code> and <code>right</code> on columns labeled <code>col1</code> and <code>col2</code> .
<code>df.pivot_table(index, columns, values=None, aggfunc='mean')</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 Series 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()/grouped.min()/grouped.max()</code>	Return a Series/DataFrame containing mean/min/max 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>

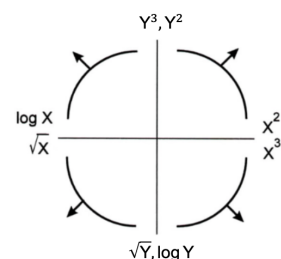
Function	Description
<code>s.str.len()</code>	Returns a Series containing length of each string
<code>s.str.lower()/s.str.upper()</code>	Returns a Series containing lowercase/uppercase version of each string
<code>s.str.replace(pat, repl)</code>	Returns a Series after replacing occurrences of substrings matching regular expression <code>pat</code> with string <code>repl</code>
<code>s.str.contains(pat)</code>	Returns a boolean Series indicating whether a substring matching the regular expression <code>pat</code> is contained in each string
<code>s.str.extract(pat)</code>	Returns a Series of the first subsequence of each string that matches the regular expression <code>pat</code> . If <code>pat</code> contains one group, then only the substring matching the group is extracted

## Visualization

Matplotlib: `x` and `y` are sequences of values.

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>

Tukey-Mosteller Buge Diagram.



Seaborn: `x` and `y` are column names in a DataFrame `data`.

Function	Description
<code>sns.countplot(data, x)</code>	Create a barplot of value counts of variable <code>x</code> from <code>data</code>
<code>sns.histplot(data, x, kde=False)</code> <code>sns.displot(x, data, rug = True, kde = True)</code>	Creates a histogram of <code>x</code> from <code>data</code> ; optionally overlay a kernel density estimator. <code>displot</code> is similar but can optionally overlay a rug plot.
<code>sns.boxplot(data, x=None, y)</code> <code>sns.violinplot(data, x=None, y)</code>	Create a boxplot of <code>y</code> , optionally factoring by categorical <code>x</code> , from <code>data</code> . <code>violinplot</code> is similar but also draws a kernel density estimator of <code>y</code> .
<code>sns.scatterplot(data, x, y)</code>	Create a scatterplot of <code>x</code> versus <code>y</code> from <code>data</code>
<code>sns.lmplot(x, y, data, 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(x, y, data, kind)</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

List of all metacharacters: `.` `^` `$` `*` `+` `?` `[]` `\` `|` `()` `{}` `}`

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
<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>

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
$L_1$ loss	$L_1(y, \hat{y}) =  y - \hat{y} $	Correlation $r$	$r = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{\sigma_x} \frac{y_i - \bar{y}}{\sigma_y}$
$L_2$ loss	$L_2(y, \hat{y}) = (y - \hat{y})^2$	Linear regression prediction of $y$	$\hat{y} = a + bx$
Empirical risk with loss $L$	$R(\theta) = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{y}_i)$	Least squares linear regression, slope $\hat{b}$	$\hat{b} = r \frac{\sigma_y}{\sigma_x}$

Least squares linear regression, intercept  $\hat{a}$   $\hat{a} = \bar{y} - \hat{b}\bar{x}$

Ordinary Least Squares

Multiple Linear Regression Model:  $\hat{\mathbb{Y}} = \mathbb{X}\theta$  with design matrix  $\mathbb{X}$ , response vector  $\mathbb{Y}$ , and predicted vector  $\hat{\mathbb{Y}}$ . If there are  $p$  features plus a bias/intercept, then the vector of parameters  $\theta = [\theta_0, \theta_1, \dots, \theta_p]^T \in \mathbb{R}^{p+1}$ . The vector of estimates  $\hat{\theta}$  is obtained from fitting the model to the sample  $(\mathbb{X}, \mathbb{Y})$ .

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

Scikit-Learn

Suppose `sklearn.model_selection` and `sklearn.linear_model` are both imported packages.

Package	Function(s)	Description
<code>sklearn.linear_model</code>	<code>LinearRegression()</code>	Returns an ordinary least squares Linear Regression model.
	<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>sklearn.model_selection</code>	<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.

SQL

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