# \*\*Review of tips dataset\*\*

## \*\*Table of contents\*\*

\* [Introduction](#Introduction)

\* [Setup](#Setup)

\* [What is the tips dataset?](#What-is-the-tips-dataset?)

\* [Some statistics on the dataset](#Some-statistics-on-the-dataset)

\* [Description 1](#Description-1)

\* [Skewness and kurtosis 1](#Skewness-and-kutrosis-1)

\* [Covariance](#Covariance)

\* [Correlation](#Correlation)

\* [Plotting statistics](#Plotting-statistics)

\* [Description discussion](#Description-discussion)

\* [Regression analysis: Table bill vs tips](#Regression-analysis:-Table-bill-vs-tips)

\* [Best fit and mean fit lines](#Best-fit-and-mean-fit-lines)

\* [Regression discussion](#Regression-discussion)

\* [Some other analysis](#Some-other-analysis)

\* [Time of dinner](#time-of-dinner)

\* [Smoker](#Smoker)

\* [Description 2](#Description-2)

\* [Covariance and correlation](#Covariance-and-correlation)

\* [Skewness and kurtosis 2](#Skewness-and-kurtosis-2)

\* [Plotting other analysis](#Plotting-other-analysis)

\* [Discussion of other analysis](#Discussion-of-other-analysis)

\* [Further potential analysis](#Further-potential-analysis)

\*[Bibliography](#Bibliography)

## \*\*Introduction\*\*

In this notebook, it is intended to review the tips dataset, which can be found online. To do this, a number of approaches will be taken. The first step will be to review the data using standard descriptive statistics, such as mean, mode, median, quartile ranges, standard deviation, etc. Other information will also be gleaned, such as the skewness, and kurtosis of the data (this will be explained in the relevant section), as well as correlation and covariance between certain data sets.

A second step will be to conduct some analysis between the fields of total-bill and tip, to determine a relationship. This analysis will also include both a best fit, and a mean fit lines. These lines can be used for various reasons, and will be discussed in the regression analysis section.

The third section, two subsets of the data will be extricated from the data, and compared to each other. The two subsets of the data chosen for this analysis are the time that the diner at the location, as well as whether the diner was a smoker or not. Similar analysis will conducted with regards to these two data sets, as was conducted on the overall dataset. A discussion of the results can be found in the discussion of further analysis section.

A final section covers other potential analysis that could be undertaken, using the same dataset.

The tips dataset that was used in the below was sourced from [here](http://<https://raw.githubusercontent.com/mwaskom/seaborn-data/master/tips.csv>). Note the data here is the raw form, to make it easier to import into our code.

\*Note: in this text, a number of ideas, or explanations may be mentioned. These have all been sourced either from the works listed in the bibliography, or from the author’s own ideas. For the sake of legibility, unless a direct quote, comparison, or specific suggestion is used, it will not be cited in the standard academic (Harvard) manner.

### Setup

Before running any analysis of the data, it is necessary to import the data, and the libraries that we will be using. In this case, it is intended to import the following libraries:

- \*\*Numpy\*\*: To run some functions for analysis.

- \*\*Pandas\*\*: The dataset will be held in various pandas dataframes. The pandas library will also be used to conduct various analytical exercises.

- \*\*Seaborn\*\*: Will be used for various plotting functionality.

- \*\*Matplotlib.pyplot\*\*: For use in some graphing functionality.

Additionally, throughout this notebook, the main dataframe that will hold the dataset will be called \*tips\_ds\*. The below code is used to import the libraries and dataset.

It should be noted that the above code does not have any outputs. The confirmation that the code was successfully imported can be found at the end of [What is the tips dataset?](#What-is-the-tips-dataset?) (below).

## \*\*What is the tips dataset?\*\*

The tips dataset is a set of data that is often used to teach analytical skills, especially around the segregation of data, and analysis of a group, or a whole set. Additionally, it could be used for extrapolation of data, where the data in question is not included in the actual dataset. The data forms part of a tutorial for the Python Seaborn package, and is used to demonstrate how to conduct plotting of data in Python.

The tips dataset was collected in a restaurant. (Jain, 2017) suggests that the data was collected over a 2.5 month period, in a branch of a chain restaurant in California, USA. The data was collected by a server (waiter/waitress), and recorded each table they served, and included a variety of information regarding the table.

This information that was recorded fell into 7 different columns. These were:

- \*\*total-bill\*\*: the total bill for each meal (numerical values to two decimal places).

- \*\*tip\*\*: the tip received for each meal (numerical values to two decimal places).

- \*\*sex\*\*: gender of the table (either male or female. Doesn’t note if this was the gender of the payer, but would be safe to assume so).

- \*\*smoker\*\*: whether the table being served was in a smoking or non-smoking section (this was California in the 1990s after all)

- \*\*day\*\*: day of the week that the meal was being served. Covers only Thursday to Sunday.

- \*\*time\*\*: the time of day that the service was being conducted. Broken down as either lunch or dinner service.

- \*\*size\*\*: the number of people in the party (whole numerical value).

In total the collector made 244 entries into the dataset.

All the above data can be confirmed by Python.

\*\*\*

## \*\*Some statistics on the dataset\*\*

In order to be able to work with the dataset, and draw any conclusions from the data, it is important to determine some of the dataset’s properties. In order to do this, we will extract basic measures, more commonly known as descriptive statistics. These statistics can then be used as a guide to further analysis, and comparison.

#### Description 1

The initial set of descriptive statistics are the mean, mode, and median of the data, as well as standard deviation, quartiles, and min and max values. Luckily, pandas has the ability to provide the values for most of these statistics, using the \*describe\* function. However, while this is useful, it is also necessary to understand what the terms provided by the \*describe\* function actually mean.

- \*\*Mean\*\*: Also known as the simple average, is the sum of all the values divided by the number of values being summed.

- \*\*Standard deviation\*\*: A measure of how are a number is from the mean. In a perfectly normal distribution, ~68% of all values would be within 1 standard deviation of the mean. Represented in the \*describe\* output as \*std\*.

- \*\*min\*\*: The lowest value within the dataset.

- \*\*25%\*\*: The value for the 25th percentile. In other words, 25% of all the values in the dataset are below this value.

- \*\*50%\*\*: The value for the 50% percentile. In other words, 50% of all the values in the dataset are below this value. This value is often called the median value.

- \*\*75%\*\*: The value for the 75th percentile. In other words, 75% of all the values in the dataset are below this value.

- \*\*Max\*\*: The largest value in the dataset.

The values for the tips dataset are retrieved as follows:

As we can see, certain measures from the descriptive statistics such as mode, skewness, and kurtosis are missing from the \*describe\* functionality. These statistics also give rise to important information regarding the dataset. The [Skewness and kurtosis](#Skewness-and-kutrosis-1) will be dealt with in their respective section. This leaves the mode to be ascertained.

The mode of a set of numbers, is the number that appears most often within that set. It is not necessarily close to the median or mean of a dataset, and can be well outside the 1st standard deviation of the data (this can happen when the distribution of the data is skewed). If no number appears most often (i.e. no number appears more often than any other number), then there is no mode for the data. Additionally, unlike the mean, and median statistics, which need to have numbers to work, the \*mode\* function can work on any set of data.

Using pandas, the mode can be determined using the \*mode\* function.

The \*describe\* function only returns data for numerical fields, while the \*mode\* function included information on the mode of the columns that contained string. The \*describe\* function can be used to provide some descriptive statistics on the data in the non-numeical columns. This is done by the inclusion of the \*include=[np.object]\* argument in the \*describe\* function. In addition to the statistics returned in the original extract, the following statistics are provided:

- \*\*Unique\*\*: The number of unique values to be found in the column.

- \*\*Top\*\*: The value returned most frequently in the column. Should return the smae value as the \*mode\* function.

- \*\*Freq\*\*: The frequency of the most commonly returned value from top/mode.

Using this \*describe(include=[np.object])\* function, the following values are returned for the tips dataset.

#### Skewness and kurtosis 1

The skewness and kurtosis are interrelated terms that are used to describe the nature of the distribution of the dataset, and how it differs from a normally distributed dataset. The definition of these terms are:

- \*\*Skewness\*\*: the direction and amount of asymmetry of the dataset about its mean. If the absolute value of the skewness is:

- greater than 1, the data is highly skewed and the distribution is very asymmetric.

- greater than 0.5 and less than 1, the data is moderately skewed, and the distribution is somewhat asymmetric.

- greater than 0, and less than 0.5, the distribution of the data is approximately symmetric.

- equal to 0, the data is normally distributed, and symmetric.

The sign of the skewness (i.e. positive or negative) also determines the skewness. Negative skewness indicates that the distribution is skewed to the left, the mean being less than the median, which is less than the mode. Positive values of skewness indicate the opposite, with the distribution being skewed to the right, and the mode being less than the median, which is less than the mean.

- \*\*Kurtosis\*\*: the kurtosis of a dataset indicates the sharpness, or flatness, of the peak of the data (around the mode, or mean, depending on the skewness).

The kurtosis is measure against the normal distribution, which has a kurtosis of 0. If the kurtosis is negative, then distribution of the data has a smaller standard deviation, as more values are grouped near the mean. This gives the distribution a sharper, and higher peak, and narrower body. A positive kurtosis indicates that there is less grouping around the mean, and indicating that the distribution has a larger standard deviation. This also gives the distribution a flatter, and lower peak, and a wider body.

In pandas, the skewness and kurtosis of a dataset can be ascertained using the \*skew\* and \*kurt\* functions. These functions return values for each numeric column within the data set.

#### Correlation

Correlation is a statistic that can be used to measure how well two sets of data correspond to each other. Weisstein (2019) defines correlation as \*“the degree to which two or more quantities are linearly associated.\* As such, a correlation coefficient can be calculated that shows the relationship between the two sets of variables, as well strength of the relationship.

In correlation analysis, positive values show that the two sets of data are positively correlated (i.e. as one value rises or falls, so does the other). Conversely, negative values indicate that the two data sets are negatively or inversely correlated (i.e. as one value rises, or falls, the other falls, or rises). A zero value indicates that there is no relationship between the two sets of data. The strength of the relationship is indicated be the actual value of the correlation coefficient. An absolute value above 0.5 is considered a strong correlation, and above 0.75 is a very strong correlation. A value of -1 or 1 means that the two sets of data are perfectly correlated (i.e. either perfectly positive or perfectly negative correlation).

In pandas dataframes, the \*corr\* function can be used to ascertain the correlation between numeric sets of data.

#### Plotting statistics

#### Covariance

#### Correlation

#### Description discussion

## \*\*Regression analysis: Table bill vs tips\*\*

#### Best and mean fit lines

#### Regression discussion

## \*\*Some other analysis\*\*

### Time of dinner

### Smoker

#### Description 2

#### Covariance and correlation

#### Skewness and kurtosis 2

#### Plotting other analysis

#### Discussion of other analysis

## \*\*Further potential analysis\*\*

## \*\*Bibliography\*\*

# Bibliography